Supplementary Information: Decoupling Visualization and Testing when Presenting Confidence Intervals

David A. Armstrong II
Professor, Department of Political Science, Western University, London, Ontario, Canada
e-mail: dave.armstrong@uwo.ca

William Poirier
Ph.D. Student, Department of Political Science, Western University, London, Ontario, Canada

Data Availability

Replication code and data for this article have been published in the Political Analysis Dataverse at https://doi.org/10.7910/DVN/GFLSLH (Armstrong II and Poirier 2024).

Appenbdix 1 Pairwise Tests vs Confidence Intervals

To be clear about the problem, consider the difference between two estimates: $b_1 - b_2$, assuming that $b_1 > b_2$. Since $b_1 > b_2$, the t-statistic below will be positive and we evaluate the t-statistic of the difference relative to its critical value t_{crit}. The difference is significant if

$$t_{\text{crit}} < \frac{b_1 - b_2}{\sqrt{V(b_1) + V(b_2) - 2V(b_1, b_2)}} \tag{1}$$

Conducting the same test with confidence intervals does something similar, but not the same. If we assume that non-overlapping intervals indicate a significant difference, we want to know whether $b_1 - t_{\text{crit}}\sqrt{V(b_1)}$ (the lower confidence bound for β_1) is bigger than $b_2 + t_{\text{crit}}\sqrt{V(b_2)}$ (the upper confidence bound for β_2). This suggests:

$$t_{\text{crit}} < \frac{b_1 - b_2}{\sqrt{V(b_1)} + \sqrt{V(b_2)}} \tag{2}$$

These tests will only produce the same test statistic when $V(b_1, b_2) = -\sqrt{v(b_1)}\sqrt{v(b_2)}$. That said, they will both produce the same result with respect to statistical significance more often than that.

The overlap in confidence intervals of statistically different estimates can be large if the correlation between the two estimates is high. For example, consider the following situation where b_1 and b_2 are two estimates whose difference is of interest. We set b_1 to zero and then vary b_2 over the range [0,5]. Further, assume that $V(b_1) = V(b_2) = 1$ and that we allow $V(b_1, b_2)$, the covariance between the two estimates to range from [-.95,.95]. Note, in this situation since both variances are 1, the covariance and the correlation are the same. The confidence interval for b_1 will be roughly $(-1.96, 1.96)$ We can then calculate the confidence interval for b_2 which will be the same as for b_1 when $b_2 = 0$ all the way to $(3.04, 6.96)$ when $b_2 = 5$. To get the percentage overlap, we first subtract the lower bound of b_2 from the upper bound of b_1. If this difference is positive, it means there is some amount of overlap in the intervals. We then divide that difference by the length of the confidence interval for b_1. We can test for significance using the formula in equation 1. If we subset the results to only those cases where there is a significant difference, we can plot the greatest percentage of overlap among significant differences as a function of the correlation between the estimates in Figure 1. Figure 1 shows that there can be more than 80% overlap if the estimates are *very* highly correlated ($r = .95$).

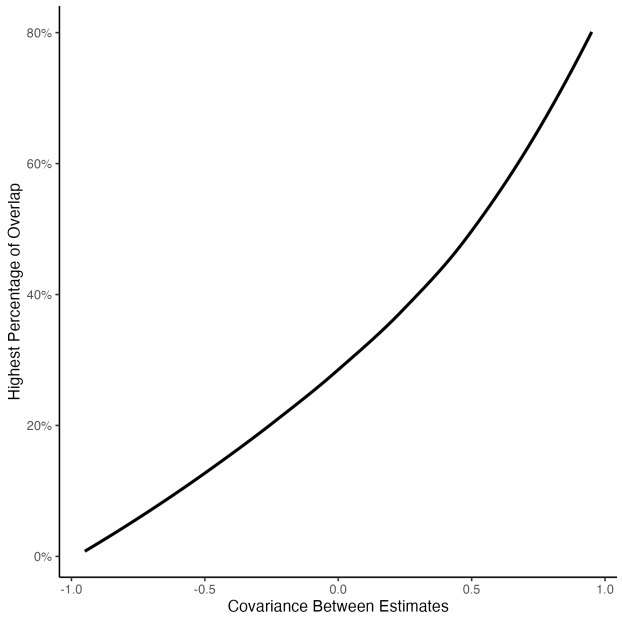


Figure 1: Maximal Percentage of Overlap of 95% Confidence Intervals for Statistically Different Estimates by Correlation

For a more moderate correlation of 0.5, the overlap of the second interval with the first is just under 50%. This example may not be indicative of all real world scenarios because the estimates often have different variances, but it does suggest that without knowing the covariance between the estimates, even what seem to be safe inferences using the overlaps in confidence intervals may be erroneous.

Appenbdix 2 Similarities to the Reference Category Problem

The so-called reference category problem exists when a set of estimates, we'll call them $g_2, \ldots, g_m$ represent regression coefficients for the dummy variables created from a categorical variable, where here we assume the first level is the reference category resulting in the identifying constraint $g_1 = 0$. The estimates, $g_2, \ldots, g_m$ allow us to directly test the difference between the reference group and each of the non-reference groups. What happens if we want to test the null hypothesis $\gamma_2 = \gamma_3$, for which $g_2 - g_3$ is an estimate. To do this, we would need to make the corresponding t-statistic:

$$t = \frac{g_3 - g_2}{\sqrt{\text{var}(g_2) + \text{var}(g_3) - 2\text{cov}(g_2, g_3)}} \tag{3}$$

The variances in the equation are generally easily obtained from the regression output by squaring the standard errors. While it is trivial to obtain the covariance of the two estimates from most statistical software, those values are not usually reported. The result is that users are left without the relevant information to conduct all relevant pairwise tests. This is where the similarity arises between visual testing and the reference category problem. There are many solutions to the reference category problem, many of them visual in nature.

One person commenting on the manuscript wondered whether quasi-variances (Firth 2003; Firth and De Menezes 2004) and particularly the quasi-confidence intervals that get produced therefrom might solve this problem? Essentially, to identify for each estimate, b_i (including the reference category) a quasi-variance, call it q_i^2 such that the relationship below holds as closely as possible:

$$\frac{b_j - b_i}{\sqrt{\text{var}(b_i) + \text{var}(b_i) - \text{cov}(b_i, b_j)}} \approx \frac{b_j - b_i}{\sqrt{q_i^2 + q_j^2}} \tag{4}$$

In terms of visual testing, this puts us in no better and often a slightly worse position in terms of visual inference to our initial result. To do tests using the quasi-variances, we would make a diagonal matrix of quasi-variances. This would be akin to the situation where we had uncorrelated estimates. We know that even in this case, the 95% intervals are often not sufficient to perform a visual test at the 5% level. In addition, except in cases with only three categories, the quasi-variances are prone to error - they do not perfectly capture all the variance and covariance information required. As such, we are essentially compounding the error in the quasi-variances with the error in the visual testing algorithm.

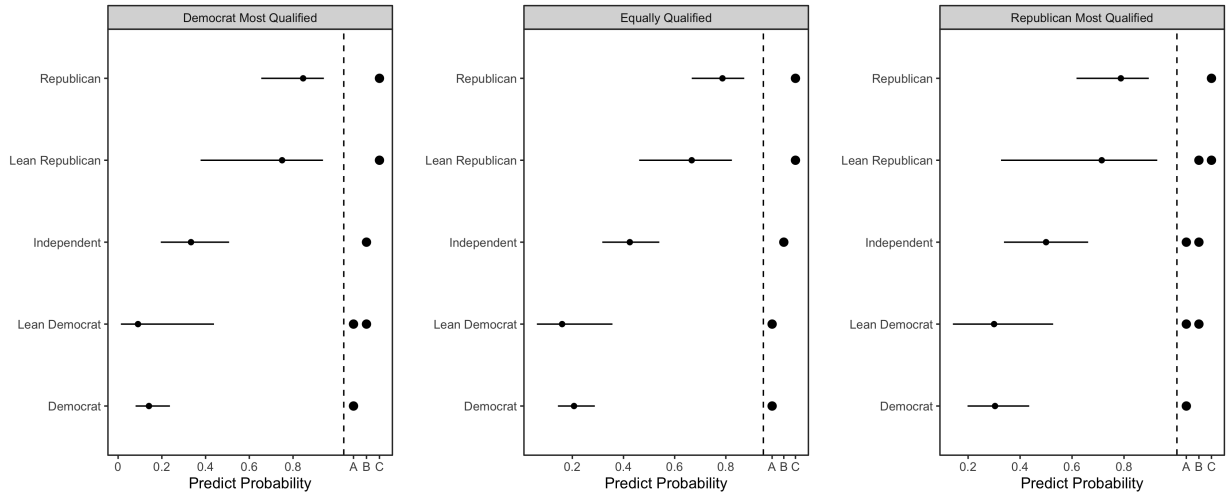


Figure 2: Iyengar and Westwood (2014)'s Predicted Probabilities for Partisan Winner Selection (CLD)

The literature on the reference category problem proposes several useful solutions, many of which are discussed in Armstrong II (2013) and Andersen and Armstrong II (2022). One such display is called a compact letter display (CLD) (Piepho 2004). This identifies groups of estimates that are not different from each other by assigning them the same "letter". Figure 2 is a CLD that corresponds to Figure 2 in the main article. This does give users the ability to make valid inferences about any pairwise comparison they want within each panel. However, this is a display that is not familiar to most political scientists and can get confusing if the patterns of differences are somewhat complex. Further, while it is relatively easy to make these displays in R, it would be more complicated in Stata and SPSS. Thus, we think our solution is superior because it relies on a visual display that is already familiar to most researchers in the field and is trivial to produce in any statistical software.

Appenbdix 3 84% Confidence Intervals

The most popular solution to the visual testing problem is arguably using 84% confidence intervals (Goldstein and Healy 1995; Payton, Greenstone, and Schenker 2003; Tukey 1991). The idea is that two 84% confidence intervals for means will overlap roughly 95% of the time under the null hypothesis. However, this only works when the ratio of standard errors for the estimates being compared is roughly 1 and the samples are independent. While this is a good start, it does not mean that 84% confidence intervals will work in general and particularly not in regression contexts where estimates are rarely independent.

If we take non-overlapping intervals to indicate statistical significance, we can calculate the type I error rate for any pair of intervals by imposing the null hypothesis condition and calculating one minus the probability that the intervals overlap. We do this for 84% confidence intervals for all combinations of $\theta = \{1, 2, 3\}$ and

$\rho = \{-.9, -.6, -.3, 0, .3, .6, .9\}$.[1] Figure 3 shows the results. In all cases, increasing the correlation between the estimates reduces the type I error rate. When the ratio of standard errors is 1, the 84% interval works if the estimates are independent. If estimates are negatively correlated, the type I error rate is higher than desired and if they are positively correlated the type I error rate is lower than desired. As the ratio of standard deviations increases, the same general pattern holds, but the correlation between the estimates that produces the right type I error rate increases. This shows that the estimates could be quite far off in real-world applications where the variances of the estimates could vary quite a lot and the correlations between estimates could be relatively far from zero. While 84% confidence intervals may work in some cases, they are not a general solution.

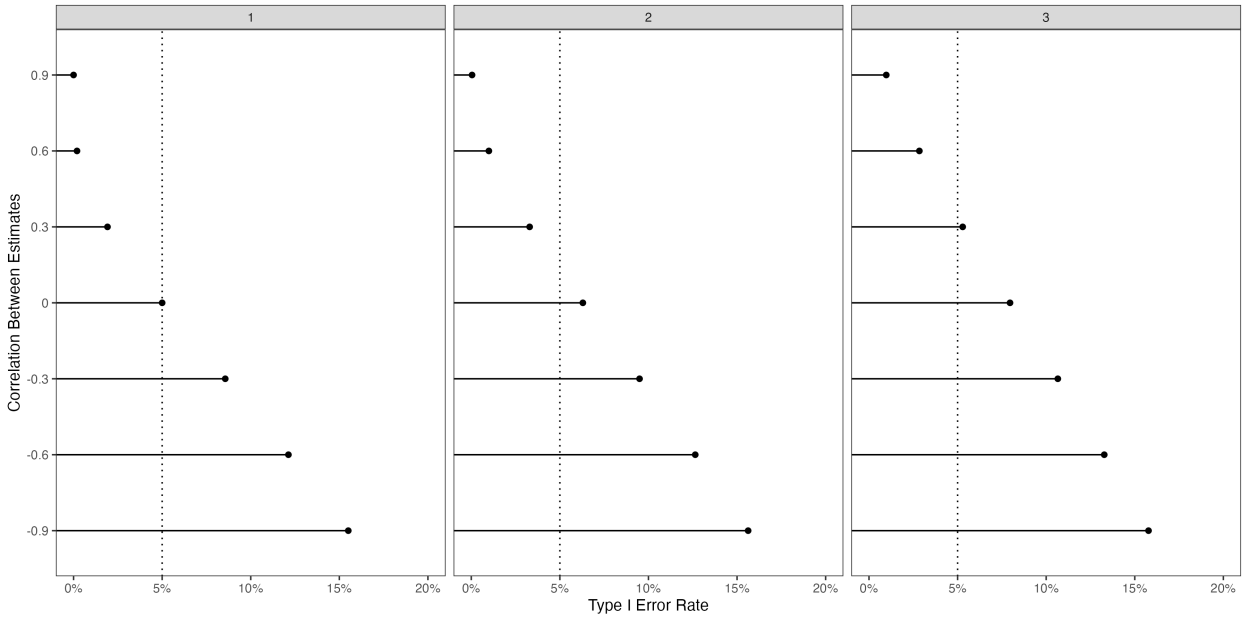


Figure 3: Type I Error Rates for 84% Confidence Intervals

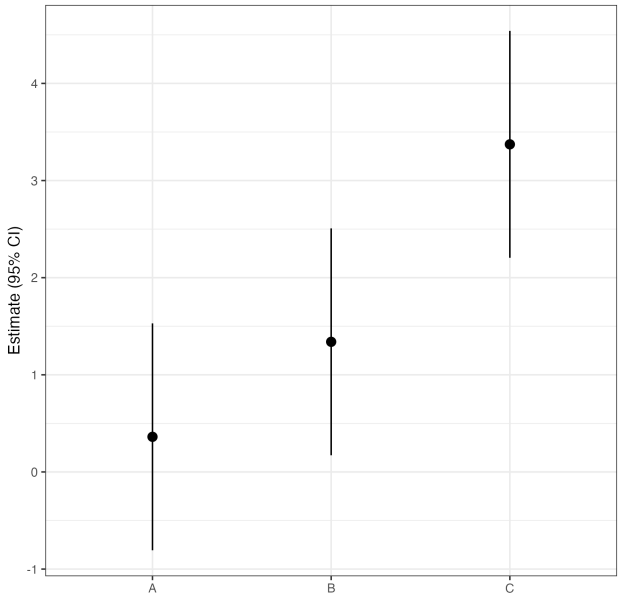
Further, even in situations where the 84% intervals *should* work, they do not always agree with the pairwise test. Using the simulation above for $\theta = 1$ and $\rho = 0$ (the situation where the overlap in 84% intervals has a Type I error rate of 5%), we find that out of 10,000 iterations, the 84% interval and the pairwise test agree a vast majority of the time, but there are 18 times (0.2%) where the 84% intervals overlap and the pairwise difference is statistically significant. The magnitude of the problem here is certainly small, but it does suggest that even in the most optimistic case, the pairwise test and 84% intervals will not *always* produce the same result.

Appenbdix 4 Choosing the Appropriate Inferential Confidence Level

Often times, a range of inferential confidence levels will all produce the same result with respect to correspondence between (non-)overlaps and test results. When this happens, the user must pick a level. To highlight the options, we use a hypothetical example. Imagine three estimates; we label them `A`, `B`, and `C`. Further, `A` and `B` are not statistically different from each other, while `B` and `C` as well as `A` and `C` are statistically different from each other. Figure 4 shows the 95% confidence intervals for the estimates. Notice that the intervals for `A` and `B` overlap as do the intervals for `B` and `C` despite the fact that the estimate for `B` is statistically different from the estimate for `C` at the 5% level. Also, we see that `A` is not statistically different from zero but both `B` and `C` are.

1. To remind, θ is the ratio of standard errors and ρ is the correlation of the estimates.

Figure 4: 95% Confidence Intervals for Hypothetical Data



Our procedure indicates that all confidence levels between 59.1% and 91.3% will perfectly represent all six tests - the three pairwise tests and the three tests of the parameter estimates relative to zero. The question we have to consider is how to choose a value in this range. There are several possibilities.

One reviewer suggested that the best level might be the one closest to the nominal rate of the test. Here, we would choose the level closest to 95%, so 91.3% would be what we would use. This as closely as possible preserves the lengths of the confidence intervals while also making them compatible with the visual tests being done. Choosing a level closer to 95% will tend to accentuate the overlapping intervals, making it easier to identify differences that are not significant.[2] Choosing the highest acceptable level will generate a plot where the intervals for the closest significantly different estimates will be separated by a very small distance, making it difficult to see whether the intervals are overlapping or not.

The upper left-hand corner of Figure 5 shows the result when choosing the 91.3% intervals. The lighter orange polygon depicts the overlap of the intervals for estimates `A` and `B`. The larger this overlap is, the easier it is to tell that estimates `A` and `B` are not statistically different from each other. The darker blue polygon (it looks like a line here) highlights the distance between the upper bound of the interval for `B` and the lower bound for `C`. The larger this polygon, the easier it is to see that the intervals for `A` and `B` do not overlap and thus are statistically different from each other. There is also the difference between `A` and `C` to consider. However, these estimates and their corresponding intervals are much farther apart than those for `B` and `C` so if we can tell whether `B` and `C` are statistically different from each other, then it should be easier to do the same for `A` and `C`.

The converse of choosing a high (or the highest) level would be to choose the smallest level. This would generate the most compact set of confidence intervals that are still maximally consistent with the visual tests. This does exactly the opposite – accentuating the distance between the two closest significantly different estimates, but making it somewhat difficult to discern whether the intervals overlap for the most distant statistically insignificant pair of estimates. The upper right-hand panel of Figure 5 shows this result. The darker blue polygon in this display is bigger and the lighter orange polygon is a line.

2. Usually, the acceptable levels will all be smaller than 95%, so the one closest to 95% will usually be the largest acceptable value.

If we learned anything from *Goldilocks and the Three Bears*, it is that the middle is always the best. The lower left-hand panel of Figure 5 depicts this scenario. Halfway between 59.1% and 91.3% is 75.2%. Here, you can see that all three tests are pretty easy to discern. The overlap between the intervals for `A` and `B` is quite clear. The distance between the upper bound for `B` and the lower bound for `C` is even easier to identify as it is larger than the overlap for `A` and `B`.

The middle value seems like a reasonable choice. In the lower left-hand panel of Figure 5, however, you can see that the overlap between the two most distant estimates that are not significantly different from each other (`A` and `B`) and the distance between the ends of the intervals for the two closest estimates that are significantly different (`B` and `C`) are of different sizes making one test slightly easier to apprehend than the other. One slight modification to picking the middle value would be to pick a value that made the two most difficult tests – the most distant two estimates that are not statistically different and the closest two estimates that are statistically different – equally easy to apprehend, as much as possible. This is what happens in the lower right-hand panel of Figure 5. The software we have finds the inferential confidence level that makes the smallest overlap for insignificant differences and the smallest non-overlap for significant differences as close to the same size as possible.

Appenbdix 5 Using Results from a Bayesian MCMC Simulations

The results of Bayesian MCMC simulation do not have p-values or type I error rates to consider. Indeed, this is marketed as a feature of the Bayesian inferential paradigm rather than a flaw. The fact that there are no p-values to consider or null hypotheses to test does not mean that this problem no longer exists. Consider the posterior distributions for two estimates of interest, β_1 and β_2 with other model parameters collected in θ, e.g., other regression coefficients, variance estimates, etc..., with joint posterior $p(\beta_1, \beta_2, \theta|\mathbf{x}, y)$ We can define $\delta = \beta_2 - \beta_1$, the posterior distribution of which is $p(\delta, \theta|\mathbf{X}, y)$. We could then calculate the posterior probability that $\beta_1 < \beta_2$:

$$P(\beta_1 < \beta_2|\mathbf{X}, \mathbf{y}) = \int_0^\infty \int_\theta p(\delta, \theta|\mathbf{X}, y) d\delta d\theta \tag{5}$$

We could then identify whether that probability indicated a credible difference, using whatever level we choose to indicate credibility.

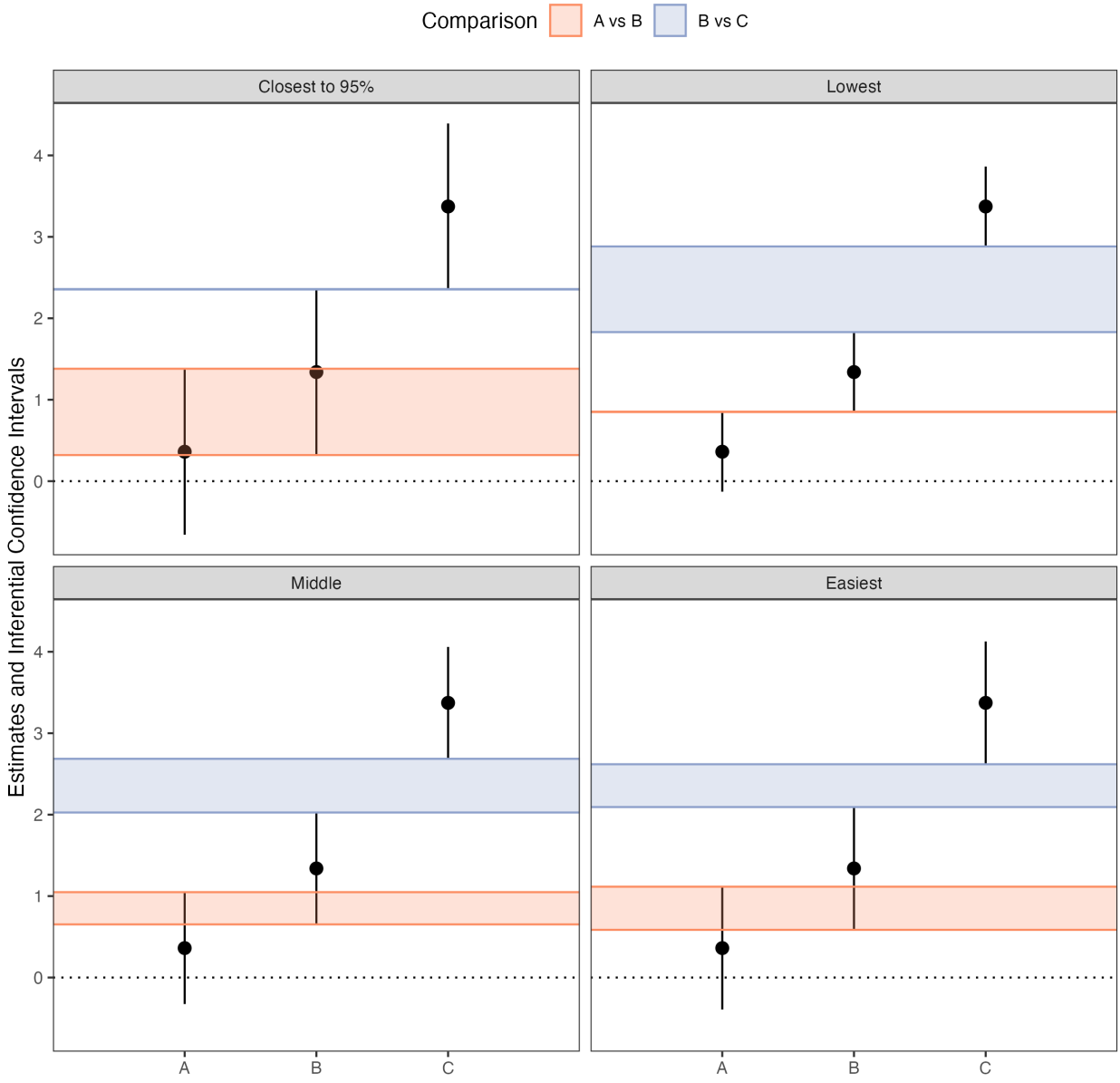
For the purposes of demonstration, consider example 12.4 from Gill (2015). In this example, time is related to several different economic indicators. The intercepts and slope coefficients relating time to the economic indicators are estimated in a hierarchical model. This produces six different intercepts and six coefficients - one of each for `DSB` (wage and salary disbursements in billions of dollars), `EMP` (employees on non-ag payrolls in thousands), `BDG` (building material sales in millions of dollars), `CAR` (auto sales in millions of dollars), `FRN` (home furnishing sales in millions of dollars) and `GMR` (general merchandise sales in millions of dollars). We will focus specifically on the slope coefficients here. The posterior summaries are presented in Table 1.

Econ Indicator	Median	MAD	$\hat{R}$	95% HDI	
DSB	0.040	0.058	1.000	-0.077	0.155
EMP	0.496	0.059	1.000	0.381	0.609
BDG	0.316	0.059	1.000	0.200	0.429
CAR	1.522	0.058	1.000	1.408	1.641
FRN	0.368	0.059	1.000	0.250	0.481
GMR	0.547	0.058	1.000	0.434	0.663

Table 1: Posterior Summary for Retail Sales Random Slope Coefficients

We could figure out how credible each pairwise difference is by calculating $Pr(\beta_k > \beta_j)$ for $k > j$ (note that in the posterior samples have been organized from smallest to largest posterior means). If we adopt the

Figure 5: Inferential Confidence Intervals for Hypothetical Data



convention that credible differences are those where at least 95% of the posterior draws for β_k are greater than those for β_j, then we could mark each pairwise difference as credible or not. We could also indicate whether the 95% highest density intervals overlap for each of those pairs. We present that information in Table 2.

You'll notice that there are two rows where the differences are credible, but the HDIs overlap. If we wanted to make a display that reflected these differences appropriately, then we could search over different highest posterior density probabilities to find one(s) where the (non-)overlaps in the HDIs correspond maximally with the credibility of the differences. We would find for this example that any probability from 0.718 to 0.877 for the HDI will have intervals that correspond perfectly with the results of our pairwise comparisons. The middle value is .795 and the value that makes the two most difficult tests easiest to apprehend is 0.816,

Indicator j	Indicator k	$\bar{\Delta}$	$\Pr(\Delta > 0)$	Δ Credible	HDI Overlap
DSB	BDG	0.276	1.000	Yes	No
DSB	FRN	0.328	1.000	Yes	No
DSB	EMP	0.456	1.000	Yes	No
DSB	GMR	0.507	1.000	Yes	No
DSB	CAR	1.483	1.000	Yes	No
BDG	FRN	0.052	0.736	No	Yes
BDG	EMP	0.180	0.987	Yes	Yes
BDG	GMR	0.231	0.997	Yes	No
BDG	CAR	1.207	1.000	Yes	No
FRN	EMP	0.128	0.940	No	Yes
FRN	GMR	0.179	0.985	Yes	Yes
FRN	CAR	1.155	1.000	Yes	No
EMP	GMR	0.051	0.732	No	Yes
EMP	CAR	1.026	1.000	Yes	No
GMR	CAR	0.975	1.000	Yes	No

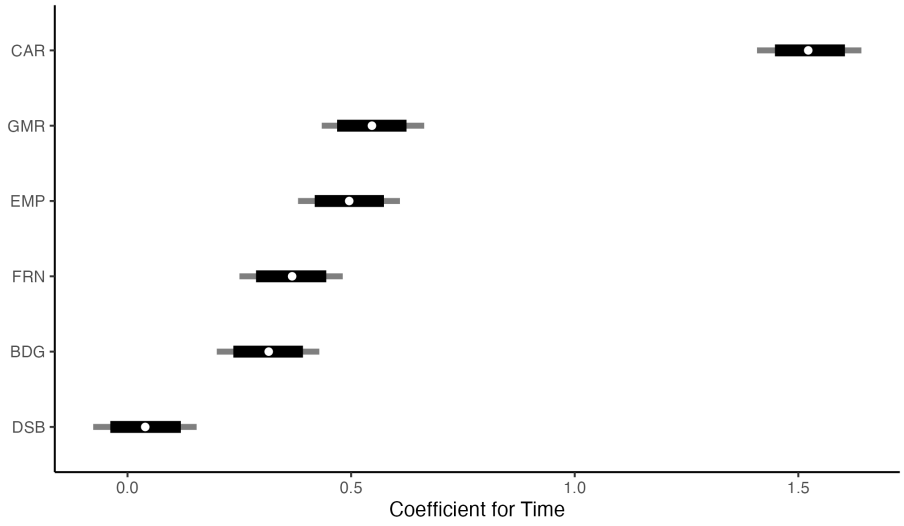
Table 2: Credibility and Overlap of HDIs for Retail Sales Random Slope Coefficients

.

so either of those values would work fine. We will use 0.816 for this demonstration.

We admit that presenting only the 81.6% HDIs changes the information presented in a meaningful way. The HDIs represent something that is somewhat more interesting than a confidence interval as they summarise the parameter of interest directly. A compromise would be to present both sets of values as in Figure 6. The light-gray lines represent the 95% highest density region and the thick black bars represent the 81.6% highest density regions where the (non-)overlaps of the intervals correspond perfectly with the credibility of the differences in estimates.

Figure 6: Inferential Posterior HDIs for Retail Sales Example



Appenbdix 6 Case Study: Muraoka and Rosas (2021)

Muraoka and Rosas (2021) consider the effect of economic status and economic inequality in the perceived placement of political parties. They show that the effect of economic inequality varies by economic status (4 groups) and by the ideological leaning of the party (Left, Centre or Right). The results derive from a Bayesian analysis. Using the idea of inferential credible intervals from Appenbdix 5, we can identify the inferential intervals that would allow users to evaluate differences among these effects.[3] We find that 76.6% credible intervals are the optimal ones for visual testing. Of the 78 tests (all pairs, and single-point tests against zero) 76.6% intervals misrepresent 5 tests (compared to 15 for the 90% intervals and 21 for the 95% intervals). Four of the missed tests are single-point tests relative to zero, which are easy to identify in the display. We do this by also presenting the 90% credible intervals whose overlap with zero is easy to evaluate. Figure 7 shows the 90% and 76.6% credible intervals.

There are a few interesting differences worth considering. Among parties on the left, there is a credible difference between those in the Bottom economic group and those in the Second-Top. This is clear from the 76.6% intervals, but not from the 90% intervals. Likewise, among parties on the Right, those in the Second-Top economic group are credibly different from those in the Second-Bottom and Bottom, but those credible differences are not apparent from the 90% credible intervals. We lose very little by using 76.6% intervals, but gain an ability to evaluate almost all pairwise differences effectively. The only test that the 76.6% intervals miss is the one between the Left Top and Left Second-Top which is significant, though the credible intervals still overlap.

These two examples (this one and the one from the main article) highlight the flexibility of our approach. It works for both Frequentist and Bayesian applications and through the use of simulation could work for any arbitrary distribution.

Appenbdix 7 Software Demonstration

Here, we demonstrate the implementation of the algorithm in R and Stata.

Appenbdix 7.1 R

The package `VizTest` contains the function that estimates the optimal confidence level. It can be installed from a GitHub repository (and eventually from CRAN).

Appenbdix 7.1.1 Gibson Replication

Below, we show the software demonstration and results for the replication of Gibson (2024).

```
remotes::install_github('davidaarmstrong/VizTest')
library(VizTest)
library(survey)

## load data
load("data/analysis/gibson_replication/gibson_dat.rda")

## make survy design object using weights
des <- svydesign(ids=~1, weight=~WEIGHT, data=gibson_dat)

## estimate model which really just calculates the survey weighted mean by wave
m <- svyglm(agree1 ~ WAVE - 1, design=des)

## find the inferential confidence levels for a two-tailed test at the 0.05 level
v <- viztest(m, test_level=.025, include_intercept = FALSE, include_zero = TRUE, level_
    increment = .001)
```

3. This is similar to the percentile intervals that Radean (2023) suggests, though again adapted for multiple comparisons.

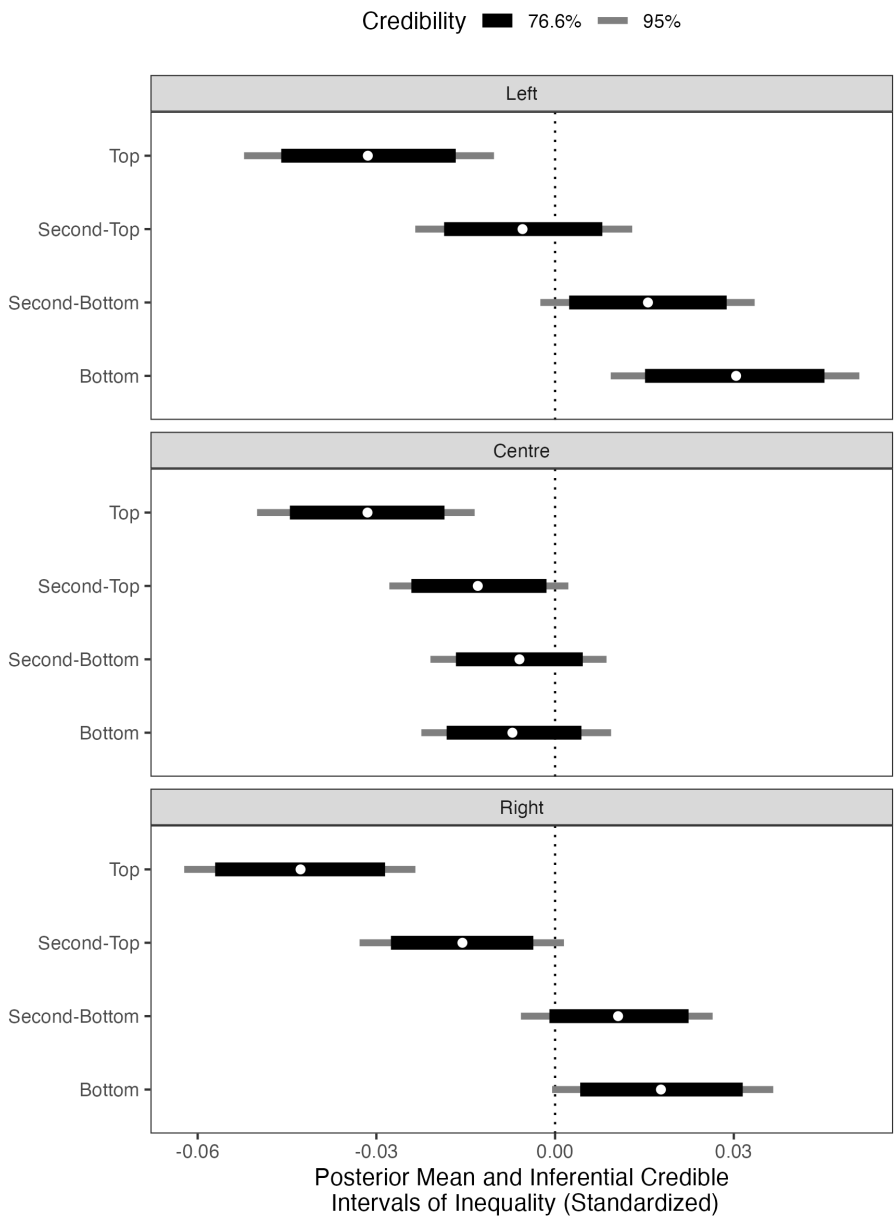


Figure 7: 90% and 76.6% Credible Intervals for the Effect of Inequality by Party Ideology and Income Group

```

## print results
v

Correspondents of PW Tests with CI Tests
  level psame pdiff         easy  method
1 0.773     1   0.8 2.730103e-07  Lowest
2 0.827     1   0.8 2.213877e-05  Middle
3 0.882     1   0.8 3.734387e-10 Highest
4 0.834     1   0.8 2.249823e-05 Easiest

All  10  tests properly represented for by CI overlaps.
```

The output here suggests that any level between 77.3% and 88.2% would work equally well. We could

use the 84% level because it is one with which users will be familiar. All ten tests - the six pairwise tests of estimates and four univariate tests of the estimates relative to zero are appropriately captured by the (non-)overlaps of the inferential confidence intervals. We plot the estimates with both levels in Figure 8.

```r
## Calculate relevant confidence intervals
ci95 <- confint(m)
ci84 <- confint(m, level = .84)

## Combine data and create labels
plot_dat <- tibble(wave = factor(levels(gibson_dat$WAVE)[1:4],
                    levels = levels(gibson_dat$WAVE)[1:4]))%>%
  bind_cols(as.data.frame(ci95) %>% setNames(c("lwr_95", "upr_95"))) %>%
  bind_cols(as.data.frame(ci84) %>% setNames(c("lwr_84", "upr_84"))) %>%
  mutate(estimate = coef(m))

## Make plot
ggplot(plot_dat, aes(x=wave, y=estimate)) +
  geom_segment(aes(y = lwr_95, yend=upr_95, colour="Original (95%)"),
                   linewidth=1.1) +
  geom_segment(aes(y = lwr_84, yend=upr_84, colour="Inferential (84%)"),
                   linewidth=3) +
  geom_point(colour="white") +
  theme_classic() +
  theme(legend.position = "top") +
  scale_colour_manual(values=c("black", "gray50"))  +
  labs(x="Survey Wave", y="Proportion Agreeing", colour="Confidence Level: ")
```

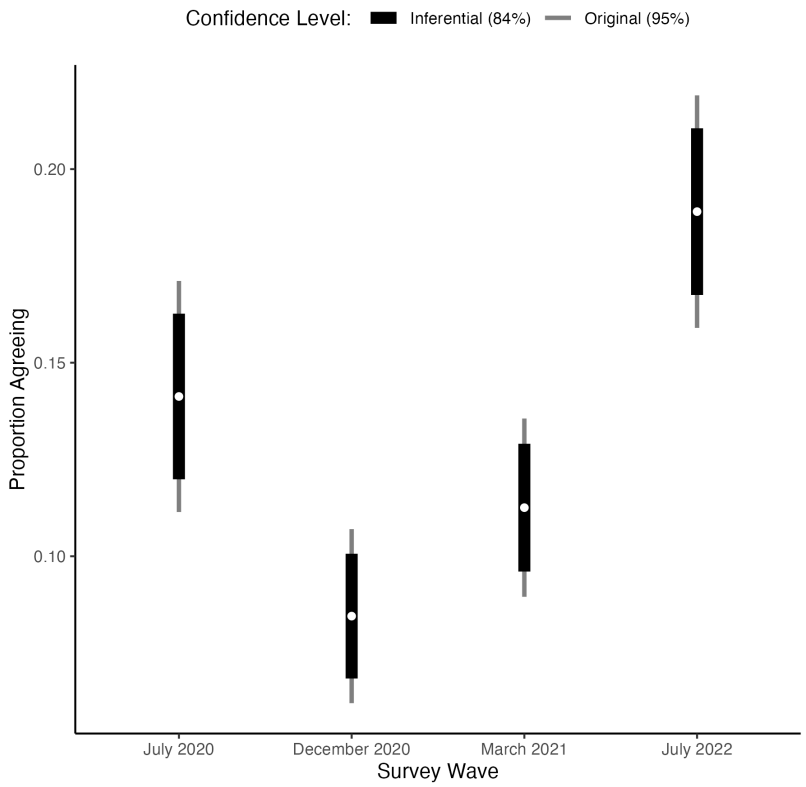


Figure 8: Gibson (2024) Replication in R

Appenbdix 7.1.2 Iyengar and Westwood Replication

Below, we show the software demonstration and results for the Iyengar and Westwood (2015) replication.

```r
## Load data
library(ggplot2)
library(tidyr)
load("data/analysis/iyengar_westwood_replication/iw_dat.rda")

## Estimate model
model<-glm(partisanSelection~participantPID2*mostQualifiedPerson,data=iw_dat[iw_dat$
    scholarship=="partisan",],family = "binomial")

## Calculate predicted probabilities for the interaction
eff<-effect(model,term="participantPID2*mostQualifiedPerson",as.table=T)

## change class of effects object to a data frame
dataeff<-as.data.frame(eff)


## get the names of the values of the most qualified person and participant party id
    variables and change them to something shorter for plotting
nms <- eff$x
nms <- nms %>%
  mutate(mostQualifiedPerson = case_when(
    mostQualifiedPerson == "Equally Qualified" ~ "EQ",
    mostQualifiedPerson == "Republican More Qualified" ~ "RMQ",
    mostQualifiedPerson == "Democrat More Qualified" ~ "DMQ"),
    participantPID2 = case_when(
      participantPID2 == "Independent" ~ "I",
      participantPID2 == "Democrat" ~ "D",
      participantPID2 == "Lean Democrat" ~ "LD",
      participantPID2 == "Republican" ~ "R",
      participantPID2 == "Lean Republican" ~ "LR")) %>%
  mutate(label = paste(participantPID2, mostQualifiedPerson, sep=":"))

## get the estimates from the effects object
b <- c(eff$fit)

## reset the names of the estimates to the labels produced above
names(b) <- nms$label

## identify the estimates that correspond with the three different "Most Qualified" options
w_eq <- grep("EQ", nms$label)
w_dmq <- grep("DMQ", nms$label)
w_rmq <- grep("RMQ", nms$label)

## Extract effects for each different "Most Qualified" option
eff_eq <- structure(list(coef=b[w_eq], vcov=vcov(eff)[w_eq, w_eq]), class="vtcustom")
eff_dmq <- structure(list(coef=b[w_dmq], vcov=vcov(eff)[w_dmq, w_dmq]), class="vtcustom")
eff_rmq <- structure(list(coef=b[w_rmq], vcov=vcov(eff)[w_rmq, w_rmq]), class="vtcustom")

## Find the inferential confidence levels for a two-tailed test at the 0.05 level for each
## different "Most Qualified" option.
vt_eq <- viztest(eff_eq, test_level = .025, level_increment = .001, include_zero=FALSE)
vt_eq

Correspondents of PW Tests with CI Tests
  level psame pdiff         easy  method
1 0.590     1   0.8 0.0005564590  Lowest
2 0.726     1   0.8 0.0535105152  Middle
3 0.863     1   0.8 0.0003973695 Highest
4 0.752     1   0.8 0.0553895827 Easiest

All  10  tests properly represented for by CI overlaps.

vt_dmq <- viztest(eff_dmq, test_level = .025, level_increment = .001, include_zero=FALSE)
vt_dmq

Correspondents of PW Tests with CI Tests
  level psame pdiff         easy  method
1 0.744     1   0.7 0.0003615722  Lowest
```

```
2 0.806     1   0.7 0.0490940109  Middle
3 0.869     1   0.7 0.0005554610 Highest
4 0.829     1   0.7 0.0539710609 Easiest

All  10  tests properly represented for by CI overlaps.

vt_rmq <- viztest(eff_rmq, test_level = .025, level_increment = .001, include_zero=FALSE)
vt_rmq

Correspondents of PW Tests with CI Tests
  level psame pdiff         easy  method
1 0.817     1   0.4 5.588237e-05  Lowest
2 0.847     1   0.4 1.019474e-02  Middle
3 0.878     1   0.4 5.030617e-04 Highest
4 0.847     1   0.4 1.019474e-02 Easiest

All  10  tests properly represented for by CI overlaps.
```

The result here suggests that for the three different Qualification treatments, each of the ten tests (six pairwise tests and four univariate tests versus zero) is captured by the (non-)overlaps of the confidence intervals within each treatment condition. A range of levels will work for each treatment condition, but anything in the range of [0.817, 0.863] will work for all conditions. Since the 84% interval is in the middle of that range, we could use it. The confidence intervals are presented in Figure 10.

```
## Calculate effects at 84% level
eff_84 <- effect(model,term="participantPID2*mostQualifiedPerson",as.table=T, se=list(level
    =.84))

## Combine original and 84% intervals
dat_all <- bind_rows(as.data.frame(eff_84) %>% mutate(interval = "Inferential (84%)"),
                     dataeff %>% mutate(interval = "Original (95%)")) %>%
  mutate(participantPID2 = factor(participantPID2,
                                  levels=c("Democrat", "Lean Democrat", "Independent", "Lean
    Republican", "Republican")),
         interval = factor(interval, levels=c("Original (95%)", "Inferential (84%)")))

rownames(dat_all) <- NULL

## Plot Results
ggplot(dat_all, aes(x=fit, xmin=lower, xmax=upper, y=participantPID2)) +
  geom_pointrange() +
  facet_grid(interval ~ mostQualifiedPerson) +
  theme_bw() +
  labs(x = "Predicted Probability of Selecting Republican", y="")
```

Appenbdix 7.1.3 Muraoka and Rosas Replication

Below, we show the software demonstration and results for the Muraoka and Rosas (2021) replication.

```
library(rstan)
load("data/raw/muraoka_rosas_replication/Left_Stan_July10.RData")
load("data/raw/muraoka_rosas_replication/Center_Stan_July10.RData")
load("data/raw/muraoka_rosas_replication/Right_Stan_July10.RData")
load("data/analysis/muraoka_rosas_replication/mr_data.rda")

## Combine estimates across ideological directions
all.eta <- cbind(left.eta, center.eta, right.eta)

## create custom testing object
all_vt <- structure(.Data = list(est = all.eta), class="vtsim")

## find inferential credible masses for the HDIs
v_all <- viztest(all_vt, test=.05, range_levels=c(.6, .9),level_increment=.001, cifun="hdi")

v_all

Correspondents of PW Tests with CI Tests
```

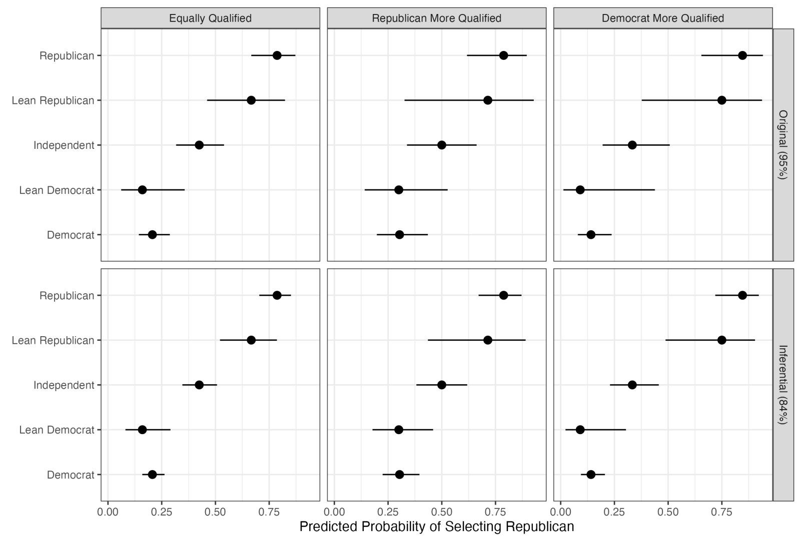


Figure 9: Iyengar and Westwood (2015) Replication in R

```
  level     psame     pdiff          easy  method
1 0.759 0.9358974 0.4615385 5.214876e-08  Lowest
2 0.769 0.9358974 0.4615385 2.891048e-07  Middle
3 0.777 0.9358974 0.4615385 2.853712e-08 Highest
4 0.768 0.9358974 0.4615385 4.332770e-07 Easiest

Missed Tests for Lowest Level (n=5 of 78)
          bigger          smaller pw_test ci_olap
28        L: Top    L: Second-Top     Sig     Yes
38 R: Second-Top             zero   Insig      No
46 C: Second-Top             zero   Insig      No
70          zero L: Second-Bottom   Insig      No
71          zero        R: Bottom   Insig      No

Missed Tests for Lowest Level (n=5 of 78)
          bigger          smaller pw_test ci_olap
28        L: Top    L: Second-Top     Sig     Yes
38 R: Second-Top             zero   Insig      No
46 C: Second-Top             zero   Insig      No
70          zero L: Second-Bottom   Insig      No
71          zero        R: Bottom   Insig      No

Missed Tests for Lowest Level (n=5 of 78)
          bigger          smaller pw_test ci_olap
28        L: Top    L: Second-Top     Sig     Yes
38 R: Second-Top             zero   Insig      No
46 C: Second-Top             zero   Insig      No
70          zero L: Second-Bottom   Insig      No
71          zero        R: Bottom   Insig      No

Missed Tests for Lowest Level (n=5 of 78)
          bigger          smaller pw_test ci_olap
28        L: Top    L: Second-Top     Sig     Yes
38 R: Second-Top             zero   Insig      No
46 C: Second-Top             zero   Insig      No
70          zero L: Second-Bottom   Insig      No
71          zero        R: Bottom   Insig      No
```

Here, we can see the first instance where some tests are not accommodated by the (non-)overlaps of the HDIs (in this case). Regardless of which level we use, they all miss the same set of tests. Four of these are relative to zero. The second-top and bottom income groups on the Right, the second-top income group in the

Centre and the second-bottom income group on the Left – all have estimates that are not credibly different from zero, but their inferential HDIs do not overlap zero. We solve this by also printing the 90% HDIs that allow us to evaluate the tests relative to zero. There is also one other difference that is not well captured. The difference between the top and second-top groups on the Left is credibly different from zero, but their inferential HDIs do overlap. That difference could be noted in the table note or in the text. The HDIs are presented in Figure **??**.

```r
## create original 95% and inferential HDIs
hdi_90 <- t(apply(all.eta, 2, \(x)HDInterval::hdi(x, credMass = .9))) %>%
  as_tibble(rownames="term") %>%
  dplyr::rename(lwr_90 = lower, upr_90 = upper)

hdi_inf <- t(apply(all.eta, 2, \(x)HDInterval::hdi(x, credMass = .768))) %>%
  as_tibble() %>%
  dplyr::rename(lwr_inf = lower, upr_inf = upper)

## combine HDIs and make appropriate variables/labels for the data
plot_dat <- bind_cols(hdi_90, hdi_inf)
plot_dat <- plot_dat %>%
  mutate(estimate = colMeans(all.eta)) %>%
  separate_wider_delim(term, ": ", names = c("Direction", "Income")) %>%
  mutate(Direction = factor(Direction, levels=c("L", "C", "R"),
                            labels = c("Left", "Centre", "Right")),
         Income = factor(Income, levels=c("Top", "Second-Top", "Second-Bottom", "Bottom")))

## Make plot
ggplot(plot_dat,
       aes(x=estimate, y=Income)) +
  geom_segment(aes(x=lwr_90, xend = upr_90, colour="Original (95%)"), linewidth=1.5) +
  geom_segment(aes(x=lwr_inf, xend=upr_inf, colour="Inferential (76.8%)"), linewidth=3) +
  geom_point(colour="white") +
  geom_vline(xintercept=0, linetype=3) +
  scale_colour_manual(values=c("black", "gray50")) +
  facet_wrap(~Direction, ncol=1) +
  theme_bw() +
  theme(panel.grid = element_blank(),
        legend.position="top") +
  labs(x="Posterior Mean and Inferential Credible\n Intervals of Inequality (Standardized)",
       y="",
       colour="HDI: ")
```

Appenbdix 7.2 Stata

You can download the optimal confidence level finding function for Stata from GitHub.

Appenbdix 7.2.1 Gibson Replication

```
net install viztest, from("https://raw.githubusercontent.com/davidaarmstrong/viztest_stata/
    main/")

* load data
use "data/analysis/gibson_replication/gibson_dat.dta", clear

* run regression
quietly reg agree1 i.WAVE [pw=WEIGHT]

* calculate effect of interest
quietly margins WAVE

* Find inferential confidence intervals
viztest, a(.025) usemargins incr(.001)

Optimal Levels:
```

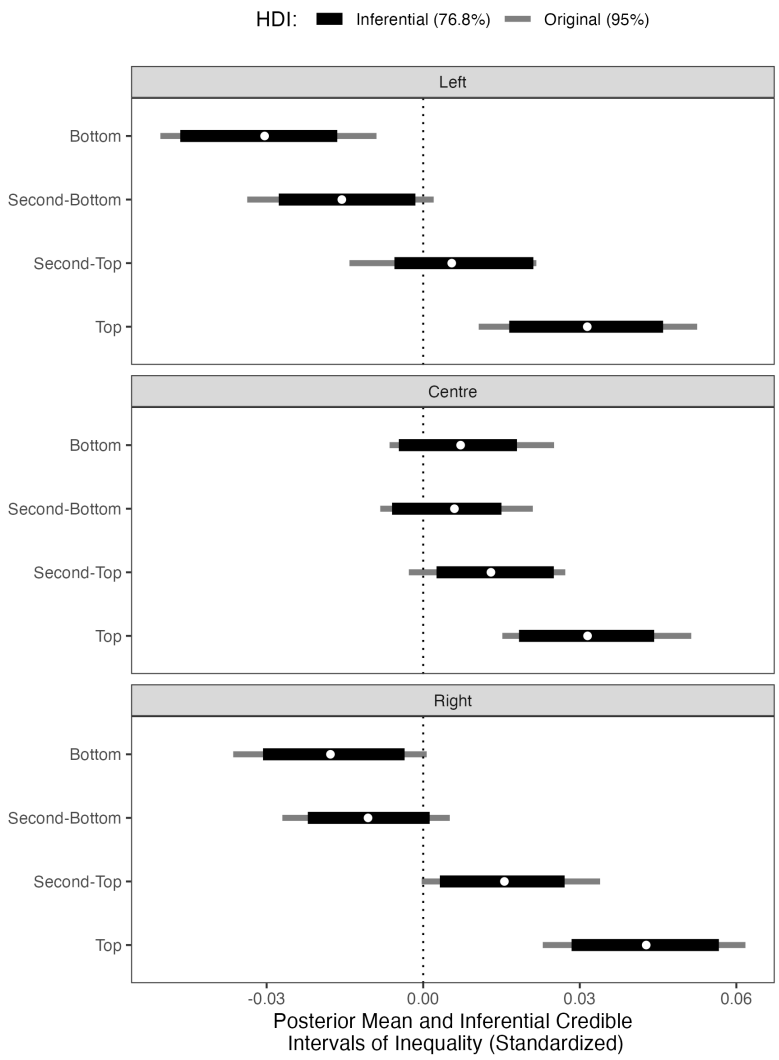


Figure 10: Muraoka and Rosas (2021) Replication in R

```

Smallest Level: .773
Middle Level: .826
Largest Level: .881
Easiest Level: .834

No missed tests!
```

The output here suggests that any level between 77.3% and 88.1% would work equally well. We could use the 84% level because it is one with which users will be familiar. All ten tests - the six pairwise tests of estimates and four univariate tests of the estimates relative to zero are appropriately captured by the (non-)overlaps of the inferential confidence intervals. Note that the results here are slightly different from the R results due to slight numerical differences in the output. The confidence intervals are in Figure 11.

```
* margins using 95% intervals
margins WAVE
* save the results table and keep the estimate and the
* confidence bounds
mat tabo = r(table)'
mat tabo = tabo[....,1], tabo[....,5], tabo[....,6]
```

```
7
8  * margins with 84% confidence intervals
9  margins WAVE, level(84)
10
11 * keep the lower and upper bounds from the estimates table
12 mat tabi = r(table)'
13 mat tabi = tabi[....,5], tabi[....,6]
14
15 * put the results together in a matrix
16 mat out = tabo, tabi
17
18 * create a new frame and change to the frame
19 frame create res
20 frame change res
21
22 * place matrix results in the new frame
23 svmat out, names(out)
24
25 * rename all the variables
26 rename out1 estimate
27 rename out2 lwr95
28 rename out3 upr95
29 rename out4 lwr84
30 rename out5 upr84
31
32 * generate a variable for the x-axis
33 gen obs = _n
34
35 * make the graph
36 twoway  (rcapsym lwr95 upr95 obs, lwidth(medium) msymbol(none) lcolor(gs8)) || ///
37     (rcapsym lwr84 upr84 obs, lwidth(vthick) msymbol(none) lcolor(black)) || ///
38     (scatter estimate obs, mcolor(white) mfcolor(white) msymbol(circle)), ///
39     xlabel(1 "July 2020" 2 "December 2020" 3 "March 2021" 4 "July 2022")
40     legend(order(2 "Inferential (84%)" 1 "Original (95%)") position(12) cols(2))
41     xtitle("Wave") ytitle("Proportion Agreeing")
42
43 * change to the default frame and drop the one
44 * created for the figure
45 frame change default
46 frame drop res
```

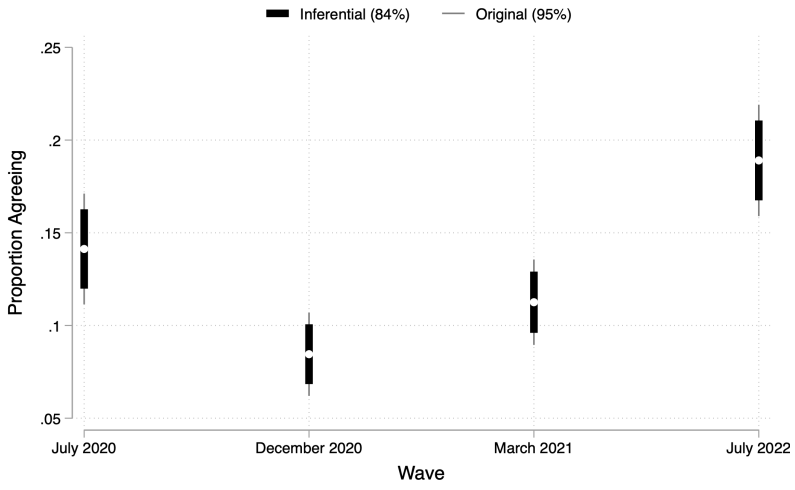


Figure 11: Gibson (2024) Replication in Stata

Appenbdix 7.2.2 Iyengar and Westwood Replication

```
use "~/Dropbox/optci/iw_dat.dta", clear

* keep required obs
keep if scholarship == "partisan"

* estimate logit with interaction
logit partisanSelection i.participantPID2##i.mostQualifiedPerson


* equally qualified
* calculate margins for party id
margins participantPID2, at(mostQualifiedPerson = 1)

* find inferential confidence level
viztest, a(.025) lev1(.5) lev2(.95) incr(.001) usemargins

Optimal Levels:

Smallest Level: .581
Middle Level: .731
Largest Level: .883
Easiest Level: .765

No missed tests!

* republican more qualified
* calculate margins for party id
margins participantPID2, at(mostQualifiedPerson = 2)

* find inferential confidence levels
viztest, a(.025) lev1(.5) lev2(.95) incr(.001) usemargins

Optimal Levels:

Smallest Level: .818
Middle Level: .843
Largest Level: .87
Easiest Level: .845

No missed tests!

* democrat more qualified
* calculate margins for party id
margins participantPID2, at(mostQualifiedPerson = 3)

* find inferential confidence level
viztest, a(.025) lev1(.5) lev2(.95) incr(.001) usemargins

Optimal Levels:

Smallest Level: .5
Middle Level: .673
Largest Level: .848
Easiest Level: .678

No missed tests!
```

The result here suggests that for the three different Qualification treatments, each of the ten tests (six pairwise tests and four univariate tests versus zero) is captured by the (non-)overlaps of the confidence intervals within each treatment condition. A range of levels will work for each treatment condition, but anything in the range of [0.818, 0.848] will work for all conditions. Since the 84% interval is in that range, we could use it. The confidence intervals are presented in Figure 12.

```
* calculate margins for plotting
quietly margins participantPID2, at(mostQualifiedPerson = (1 2 3))
```

```
3
4  * save results in tabo and keep only estimate
5  * and lower/upper confidence bounds
6  mat tabo = r(table)'
7  mat tabo = tabo[....,1], tabo[....,5], tabo[....,6]
8
9  * calculate margins with inferential confidence interval
10 quietly margins participantPID2, at(mostQualifiedPerson = (1 2 3)) level(84)
11
12 * save results in tabo and keep only lower/upper confidence bounds
13 mat tabi = r(table)'
14 mat tabi = tabi[....,5], tabi[....,6]
15
16 * put results together
17 mat out = tabo, tabi
18
19 * create a new frame and change to the frame
20 frame create res
21 frame change res
22
23 * place matrix results in the new frame
24 svmat out, names(out)
25
26 * rename all variables
27 rename out1 estimate
28 rename out2 lwr95
29 rename out3 upr95
30 rename out4 lwr84
31 rename out5 upr84
32
33 * generate most qualified person variable
34 gen mqp = .
35
36 * replace values to correspond with output from margins
37 replace mqp = 1 in 1/5
38 replace mqp = 2 in 6/10
39 replace mqp = 3 in 11/15
40
41 * define an apply levels for mqp
42 label def mqp 1 "Equally Qualified" 2 "R More Qualified" 3 "D More Qualified"
43 label val mqp mqp
44
45 * generate party id variable
46 gen pid = .
47
48 * repalce values to correspond with output from margins
49 * democrats
50 foreach i of num 2 7 12 {
51   replace pid = 1 in `i'
52 }
53 * lean democrat
54 foreach i of num 3 8 13 {
55   replace pid = 2 in `i'
56 }
57
58 * independent
59 foreach i of num 1 6 11 {
60   replace pid = 3 in `i'
61 }
62 * lean republican
63 foreach i of num 4 9 14 {
64   replace pid = 4 in `i'
65 }
66 * republican
67 foreach i of num 5 10 15 {
68   replace pid = 5 in `i'
69 }
70
```

```
* define and apply labels
label def pid 1 "D" 2 "LD" 3 "I" 4 "LR" 5 "R"
label val pid pid

* Make the graph
twoway (pcspike pid lwr95 pid upr95, lwidth(medium) lcolor(gs8)) || (pcspike pid lwr84 pid
    upr84, lwidth(vthick) lcolor(black)) || (scatter pid estimate, mcolor(white) mfcolor(
    white) msymbol(circle)), by(mqp, cols(3) compact note("")) legend(order(2 "Inferential
    (84%)" 1 "Original (95%)") position(12) cols(2)) xtitle("Predict Pr(Choose Republican)")
     ytitle("") ylabel(1 "D" 2 "LD" 3 "I" 4 "LR" 5 "R")
```

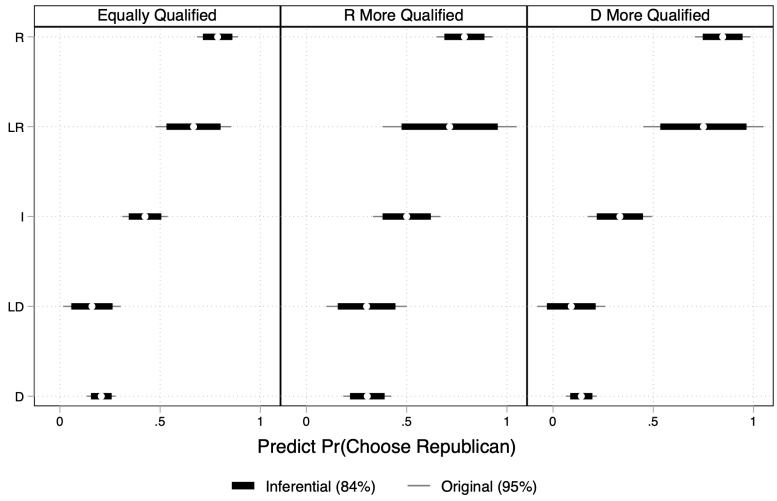


Figure 12: Iyengar and Westwood (2015) Replication in Stata

Note, the Stata results here are a bit different because the output from `margins` uses the delta method to derive standard errors for the predicted probabilities and then normal theory confidence intervals around the predicted probabilities using those standard errors.

Appenbdix 7.2.3 Muraoka and Rosas Replication

Currently, the Stata version of the software only supports Frequentist results and those from the `margins` function. As such, we do not present the Muraoka and Rosas replication in Stata.

Appenbdix 8 Cautions and Caveats

One reviewer was less sanguine about the use of this tool than we are. We want to highlight some of those thoughts here. The reviewer suggested that authors should only present confidence intervals (specifically our inferential confidence intervals, but perhaps any) for estimates where *any* pairwise comparison is reasonable. We think this is generally good practice independent of our intervention in the article. Plotting estimates on the same scale implicitly invites comparison. As researchers, we should be careful about what comparisons we invite people to make. The reviewer goes on to say that not following this advice invites substantive comparisons that may not be supported by or consistent with the literature to which the article in question contributes. Perhaps this is true. However, the terms are estimated by the model and someone may hypothesize, independently of the article's intent, about the nature of some comparison. If the model is generating estimates whose values ought not to be compared, even if due consideration is given to underlying variability and the distribution of the variables that give rise to the estimates, that seems like a modeling problem more than a presentation problem.

References

Andersen, R., and D. A. Armstrong II. 2022. *Presenting statistical results effectively.* London: Sage.

Armstrong II, D. A. 2013. factorplot: Improving Presentation of Simple Contrasts in Generalized Linear Models. *The R Journal* 5 (2): 4–15. https://doi.org/10.32614/RJ-2013-021. https://doi.org/10.32614/RJ-2013-021.

Armstrong II, D. A., and W. Poirier. 2024. *Replication Data for: Decoupling Visualization and Testing when Presenting Confidence Intervals.* V. DRAFT VERSION. https://doi.org/10.7910/DVN/GFLSLH.

Firth, D. 2003. Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology* 33:1–18.

Firth, D., and R. De Menezes. 2004. Quasi-variances. *Biometrika* 91 (1): 65–80.

Gibson, J. L. 2024. Losing legitimacy: the challenges of the dobbs ruling to conventional legitimacy theory. *American Journal of Political Science.*

Gill, J. 2015. *Bayesian methods: a social and behavioral sciences approach,* 3^{rd} *ed.* Boca Raton, FL: CRC Press.

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

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

Muraoka, T., and G. Rosas. 2021. Does economic inequality drive voters' disagreement about party placement? *American Journal of Political Science* 65 (3): 582–597.

Payton, M. E., M. H. Greenstone, and N. 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.

Piepho, H.-P. 2004. An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics* 13 (2): 456–466.

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

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