The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 24, Number 2, Spring 2017

Editor:
Justin Esarey, Rice University
justin@justinesarey.com

Associate Editors:
Randolph T. Stevenson, Rice University
stevenso@rice.edu

Editorial Assistant:
Kristin Bryant, Rice University
kab18@rice.edu

Rick K. Wilson, Rice University
rkw@rice.edu

Contents

Notes from the Editors

As *The Political Methodologist* continues its transition to a new format, the current editorial team has agreed to stay on for an extra year (two new issues) to provide sufficient time for the transition to occur. And we are pleased to present this new issue of *TPM* with great new contributions!

Two of our articles in this issue are about the proper specification and interpretation of results from parametric models: an article by Enns and Wlezien, and another by MacKinnon and Webb. Enns and Wlezien clarify how researchers can ensure that their ECMs achieve proper balance; achieving appropriate balance (i.e., an equal order of integration on both sides of the equation) is important because imbalance can create a spurious relationship between the independent and dependent variables. The authors respond in part to a recent *Political Analysis* symposium about error-correction models (or ECMs), elaborating and clarifying some aspects of imbalance that they believe the symposium left unstated or unclear.

MacKinnon and Webb write about the analysis of data with unmodeled error correlation within units, such as states. Not only do political scientists frequently analyze this kind of data, but they also frequently encounter data sets wherein the number of data points inside a unit varies substantially from unit to unit; for example, a survey conducted within the United States will almost certainly have more respondents from California than Wyoming or Montana. The authors find that wild cluster bootstrapping (or, with very few clusters, ordinary wild bootstrapping) can produce more accurate standard errors when cluster-robust standard errors or the pairs cluster bootstrap often rejects the null too often.

As part of their study, MacKinnon and Webb replicate an analysis by Burden, Canon, Mayer, and Moynihan very recently published in *Political Research Quarterly*. Their replication effort turns up some problematic aspects of one model from Burden et al., who respond to MacKinnon and Webb's analysis in this issue. In addition to acknowledging the problems with this model, Burden et al. argue that "the greater lesson from the skilled analysis of MacKinnon and Webb is to raise further doubt about whether [difference-in-difference modeling] is simply unsuitable" in settings where very few of the units (in this case, American states) actually experience a policy change of interest.

Finally, a large and interdisciplinary group of scholars led by Daniel Benjamin and Valen Johnson recently proposed that statistical significance be conventionally redefined from $p < 0.05$ to $p < 0.005$ as a means of alleviating the "replication crisis" that has swept through the social sciences. A piece in this issue by *TPM* editor Justin Esarey argues that, although this policy disadvantages junior and underfunded scholars, these disadvantages may be overcome if researchers re-focus their efforts on jointly testing multiple hypotheses from a proposed theory. Joint hypothesis tests enable tests with strict size requirements to be conducted with much greater power than the same number of separate hypothesis tests.

We hope that you enjoy the articles in this issue!

The Editors

Articles

Understanding Equation Balance in Time Series Regression

Peter K. Enns
Cornell University
peterenns@cornell.edu

Christopher Wlezien
University of Texas at Austin
wlezien@austin.utexas.edu

Political Analysis (PA) recently hosted a symposium on time series analysis that built upon De Boef and Keele's (2008) influential time series article in the *American Journal of Political Science*. Equation balance was an important point of emphasis throughout the symposium. In their classic work on the subject, Banerjee et al. (1993, 164) explain that an unbalanced equation is a regression, "in which the regressand is not the same order of integration as the regressors, or any linear combination of the regressors." The contributors to this symposium were right to emphasize the importance of equation balance, as unbalanced equations can produce serially correlated residuals (e.g., Pagan and Wickens 1989) and spurious relationships (e.g., Banerjee et al. 1993, 79).

Throughout the *PA* symposium, however, equation balance is defined and applied in different ways. Grant and Lebo (2016, 7) follow Banerjee et al.'s definition when they explain that a general error correction model (GECM)—or autoregressive distributed lag (ADL)—is balanced if cointegration is present.[1] Keele, Linn and Webb (2016*a*, 83) implicitly make this same point in their second contribution to the symposium when they cite Bannerjee et al. (1993) in their discussion of equation balance. Yet, other parts of the symposium seem to apply a stricter standard of equation balance, stating that when estimating a GECM/ADL all time series must be the same order of integration. As Grant and Lebo write in the abstract of their first article, "Time series of various orders of integration—stationary, non-stationary, explosive, near- and fractionally integrated—should not be analyzed together ... That is, without equation balance the model is misspecified and hypothesis tests and long-run-multipliers are unreliable." Keele, Linn and Webb (2016*b*, 34) similarly write, "no regression model is appropriate when the orders of integration are mixed because no long-run relationship can exist when the equation is unbalanced." Box-Steffensmeier and Helgason (2016, 2) make the point by stating, "when studying the relationship between two (or more) series, the analyst must ensure that they are of the same level of integration; that is, they have to be balanced." Although Freeman (2016) offers a more nuanced perspective on equation balance, many of the symposium contributors could be interpreted as recommending that scholars never mix orders of integration.[2] Indeed, in their concluding article, Lebo and Grant write, "One point of agreement among the papers here is that equation balance is an important and neglected topic. One cannot mix together stationary, unit-root, and fractionally integrated variables in either the GECM or the ADL" (p.79).

It is possible that these authors did not mean for these quotes to be taken literally. However, we both have recently been asked to review articles that have used these quotes to justify analytic decisions with time series data.[3] Thus, we think the claims should be reviewed carefully. This is especially the case because Grant and Lebo could be interpreted as applying these strict standards in some of their empirical applications. For example, in their discussion of Sánchez Urribarrí, Schorpp, Randazzo and Songer (2011), Grant and Lebo write, "both the UK and US models are unbalanced—each DV is stationary, and the inclusion of unit-root IVs has compromised the results" (Supplementary Materials, p.36). Researchers might take this statement to imply that including stationary and unit root variables automatically produces an unbalanced equation.

In addition to holding implications for practitioners, the strict interpretation of equation balance holds implications for the vast number of existing time series articles that employ GECM/ADL models without pre–whitening the data to ensure equal orders of integration across all series. Lebo and Grant (2016, 79) point out, for example, "FI [fractional integration] methods allow us to create a balanced equation from dissimilar data. By filtering each series by its own (p, d, q) noise model, the residuals of each can be rendered (0, 0, 0) so that you can investigate how X's deviations from its own time-dependent patterns affect Y's deviations from its own time-dependent patterns." Fortunately, existing time series analysis that does not pre-whiten the data need not be automatically dismissed. The strict interpretation of equation balance—i.e., that mixing orders of integration is always problematic with the GECM/ADL—is not accurate. As noted above, the contributors to the symposium may indeed understand this point. But based on the quotes above, we feel that it is important to clarify for practitioners that an unbalanced equation is not synonymous

[1] The GECM and ADL are the same model (e.g., Banerjee et al. 1993, De Boef and Keele 2008, Esarey 2016). However, since the two models estimate different quantities of interest (Enns, Kelly, Masaki and Wohlfarth 2016), they are often discussed as two separate models.

[2] Specifically, Freeman (2016, 50) explains, "KLW's [Keele, Linn, and Webb] claim that unbalanced equations are 'nonsensical' (16, fn.4) and GL's [Grant and Lebo] recommendation to 'set aside' unbalanced equations (7) are a bit overdrawn. Banerjee et al. (1993) and others discuss the estimation of unbalanced equations. They simply stress the need to use particular nonstandard distributions in these cases."

[3] Given the prominence of the authors as well as the *Political Analysis* journal, it is perhaps not surprising that practitioners have begun to adopt these recommendations.

with mixing orders of integration. While related, they are not the same, and while the former is always a problem the latter is not.

We begin by showing that equation balance does not necessarily require that all series have the same order of integration with the GECM/ADL. This is important because the classic examples in the literature of unbalanced equations include series of different orders of integration (see, for example, Banerjee et al. (1993, 79) and Maddala and Kim (1998, 252)). But our results are not at odds with these scholars, as their examples all assume a relationship with no dynamics. When using a GECM/ADL to model dynamic processes, even mixed orders of integration can produce balanced equations. This conclusion is consistent with Banerjee et al. (1993), who write, "The moral of the econometricians' story is the need to keep track of the orders of integration on both sides of the regression equation, *which usually means incorporating dynamics*; models that have restrictive dynamic structures are relatively likely to give misleading inferences simply for reasons of inconsistency of orders of integration" (p.192, italics ours).

We believe the *PA* symposium was not sufficiently clear that adding dynamics can solve the equation balance problem with mixed orders of integration. Thus, a key contribution of our article is to show how appropriate model specification can be used to produce equation balance and avoid inflating the rate of spurious regression – even when the model includes series with different orders of integration. Our particular focus is analysis that mixes stationary $I(0)$ and integrated $I(1)$ time series. In practice, researchers might encounter other types of time series, such as fractionally integrated, near-integrated, or explosive series. Evaluating every type of time series and the vast number of ways different orders of integration could appear in a regression model is beyond the scope of this paper. Our goal is more basic, but still important. We aim to demonstrate that there are exceptions to the claim that, "The order of integration needs to be consistent across all series in a model" (Grant and Lebo 2016, 4) and that these exceptions can hold important implications for social science research.

More specifically, we show that when data are either stationary or (first order) integrated, scenarios exist when a GECM/ADL that includes both types of series can be estimated without problem. Our simulations show that regressing an integrated variable on a stationary one (or the reverse) does not increase the risk of spurious results when modeled correctly. While this may be a simple point, we think it is a crucial one. As mentioned above, if readers interpreted the previous quotes from the *PA* symposium as defining equation balance to mean that different orders of integration cannot be mixed, most existing research that employs the ADL/GECM model would be called into question. Given the fact that *Political Analysis* is one of the most cited journals in political science and the symposium included some of the top time series practitioners in the discipline, we believe it is valuable to clarify that mixing orders of integration is not always a problem and that existing time series research is not inherently flawed. Furthermore, the one article that has responded to particular claims made in the symposium contribution did not address the symposium's definition of equation balance (Enns et al. 2016).[4] We hope our article helps clarify the concept of equation balance for those who use time series analysis.

We also illustrate the importance of our findings with an applied example—income inequality in the United States. The example illustrates how the use of pre-whitening to force variables to be of equal orders of integration (when the equation is already balanced) can be quite costly, leading researchers to fail to detect relationships.[5]

Clarifying Equation Balance

The contributors to the *PA* symposium were all correct to emphasize equation balance. Time series analysis requires a balanced equation. An unbalanced equation is mis-specified by definition, typically resulting in serially correlated residuals and an increased probability of Type I errors.[6] As noted above, Banerjee et al. (1993, 164, italics ours) explain that an unbalanced equation is a regression, "in which the regressand is not the same order of integration as the regressors, *or any linear combination of the regressors*." Our primary concern is that much of the discussion in the *PA* symposium seems to focus on the order of integration of *each* variable in the equation without acknowledging that a "linear combination of the regressors" can also produce equation balance. We worry that researchers might interpret this focus to mean that equation balance requires each series in the model to be the same order of integration.[7] Such a conclusion would be wrong. As the previous quote from Banerjee et al. (1993) indicates (also see, Maddala and Kim (1998, 251), if the regressand and the regressors are *not* the same order of integration, the equation will still be balanced if a *linear combination* of the variables is the same order of integration.

As Grant and Lebo (2016, 7) and Keele, Linn and Webb (2016*a*, 83) acknowledge, cointegration offers a useful illustration of how an equation can be balanced even when the regressand and regressors are *not* the same order of integra-

[4] Enns et al. (2016) focused on how to correctly implement and interpret the GECM.

[5] Of course, equation balance is not the only relevant consideration. Researchers must check that their model satisfies other assumptions, such as no autocorrelation in the residuals and no omitted variables.

[6] See, e.g., Banerjee et al. (1993, 164-168), Maddala and Kim (1998, 251-252), and Pagan and Wickens (1989, 1002).

[7] For example, Grant and Lebo (p.7-8) write, "Additionally, any loss of equation balance makes a cointegration test dubious so, again, if the dependent variable is $I(1)$, then the model should only include $I(1)$ independent variables."

tion.[8] Consider two integrated $I(1)$ variables, Y and X, in a standard GECM model:

$$\Delta Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \gamma_1 \Delta X_t + \beta_1 X_{t-1} + \epsilon_t. \quad (1)$$

Clearly, the equation mixes orders of integration. We have a stationary regressand (ΔY_t) and a combination of integrated (Y_{t-1}, X_{t-1}) and stationary (ΔX_t) regressors. However, if X and Y are cointegrated, the equation is still balanced. To see why, we can rewrite Equation 1 as:

$$\Delta Y_t = \alpha_0 + \alpha_1 (Y_{t-1} + \frac{\beta_1}{\alpha_1} X_{t-1}) + \gamma_1 \Delta X_t + \epsilon_t. \quad (2)$$

X and Y are cointegrated when X and Y are both integrated (of the same order) and α_1 and β_1 are non-zero (and $\alpha_1 < 0$). Because cointegration ensures that Y and X maintain an equilibrium relationship, a linear combination of these variables exists that is stationary (that is, if we regress Y on X, in levels, the residuals would be stationary).[9] As noted above, this (stationary) linear combination is captured by $(Y_{t-1} + \frac{\beta_1}{\alpha_1} X_{t-1})$. Additionally, since Y and X are both integrated of order one, ΔY and ΔX will be stationary. Thus, cointegration ensures that the equation is balanced: the regressand (ΔY_t) and either the regressors (ΔX_t) or a linear combination of the regressors ($Y_{t-1} + \frac{\beta_1}{\alpha_1} X_{t-1}$) are *all* stationary. Importantly, if we added a stationary regressor to the model, e.g., if we thought innovations in Y were also influenced by a stationary variable, the equation would still be balanced.

The fact that the GECM—which mixes stationary and integrated regressand and regressors—is appropriate when cointegration is present demonstrates that equation balance does not require the series to be the same order of integration. As we have mentioned, Grant and Lebo (2016, 7) acknowledge that a GECM is balanced if co-integration is present and Keele, Linn and Webb (2016*a*, 83) make this point in their second contribution to the symposium citing Bannerjee et al. (1993) in their discussion of equation balance. However, as noted above, we have begun to encounter research that interprets other statements in the symposium to mean that analysts can never mix orders of integration. For example, in their discussion of Volscho and Kelly (2012), Grant and Lebo write that the "data is a mix of data types (stationary and integrated), so any hypothesis tests will be based on unbalanced equations" (supplementary appendix, p.48). But is this this really the case? The above example shows that when cointegration is present, equation balance can exist even when the orders of integration are mixed.

Below, we use simulations to illustrate two seemingly less well-known scenarios when equation balance exists despite different orders of integration. Again, our goal is not to identify all cases where different orders of integration can result in equation balance. Rather, we want to show that researchers should not automatically equate different orders of integration with an unbalanced equation. Situations exist where it is completely appropriate to estimate models with different orders of integration.

Equation Balance with Mixed Orders of Integration: Simulation Results

We begin with an integrated Y and a stationary X. At first glance, estimating a relationship between these variables, which requires mixing an $I(1)$ and $I(0)$ series, might seem problematic. Grant and Lebo (2016, 4) explain, "Mixing together series of various orders of integration will mean a model is misspecified" and in econometric texts, mixing $I(1)$ and $I(0)$ series offers a classic example of an unbalanced equation (Banerjee et al. 1993, 79; Maddala and Kim 1998, 252).[10]

It is still possible to estimate the relationship between an integrated Y and a stationary X in a correctly specified and balanced equation. First, we must recognize that when Banerjee et al. (1993) (see also Mankiw and Shapiro (1986) and Maddala and Kim (1998)) state that an $I(1)$ and $I(0)$ series represent an unbalanced equation, they are modeling the equation:

$$y_t = \alpha + \beta x_{t-1} + u_t. \quad (3)$$

Equation 3 is indeed unbalanced (and thus misspecified) as the regressand is integrated and the regressor is stationary. This result does not, however, mean that we cannot consider these two series. A stationary series, X, might be related to *innovations* in an integrated series, Y. If so, we could model this process with an autoregressive distributed lag model:

$$Y_t = \alpha + \alpha_1 Y_{t-1} + \beta_1 X_t + \beta_2 X_{t-1} + \epsilon. \quad (4)$$

Much as before, this might appear to still be an unbalanced equation. We continue to mix $I(1)$ and $I(0)$ series, which seemingly violates Lebo and Grant's (2016, 71) conclusion

[8] See Murray (1994) for a discussion of cointegration.

[9] This, in fact, is the first step of the Engle-Granger two-step method of testing for cointegration.

[10] Interestingly, existing simulations show that despite being unbalanced regressions, we will *not* find evidence that unrelated $I(0)$ and $I(1)$ series are (spuriously) related in a simple bivariate regression if the $I(0)$ variable is $AR(0)$ (see, e.g., Banerjee et al. (1993, 79), Granger, Hyung and Jeon (2001, 901), and Maddala and Kim (1998, 252)). Banerjee et al. explain that the only way in which OLS can make the regression consistent and minimize the sum of squares is to drive the coefficient to zero (p.80). Our own simulations confirm that when estimating unbalanced regressions with $AR(1)$ and $I(1)$ series, both serial correlation and inflated Type I error rates emerge.

Table 1: The Percent of Spurious Regressions with an Integrated Y and Stationary X

	$\theta_x = 0$				$\theta_x = 0.5$			
$T =$	$\hat{\beta}_1$	%	$\hat{\beta}_2$	%	$\hat{\beta}_1$	%	$\hat{\beta}_2$	%
50	.0032	4.2%	-.0021	5.4%	0.0025	4.2%	-0.0035	4.3%
1,000	-.0011	4.8%	0.0003	5.0%	-0.0011	4.7%	0.0010	4.3%

Notes: $\hat{\beta}$ represents the mean coefficient estimate across 1,000 simulations. % represents the percent of the simulations for which we *incorrectly* reject the null hypothesis of no relationship between X and Y. θ_x represents the autoregressive parameter in Equation 7.

that, "One cannot mix together stationary, unit–root, and fractionally integrated variables in either the GECM or the ADL."[11] However, since Y is $I(1)$, $\alpha_1 = 1$, which means $Y_t - \alpha_1 Y_{t-1} = \Delta Y$. Thus, we can rewrite the equation as,

$$\Delta Y_t = \alpha + \beta_1 X_t + \beta_2 X_{t-1} + \epsilon. \tag{5}$$

Because Y is an integrated, $I(1)$, series, ΔY must be stationary. Thus, the regressand and regressors are all $I(0)$ series. As Banerjee et al. (1993, 169) explain, "regressions that are linear transformations of each other have identical statistical properties. What is important, therefore, is the *possibility* of transforming in such a way that the regressors are integrated of the same order as the regressand."[12] Thus, Equation 5 shows that the ADL in Equation 4 is indeed balanced. (Because the GECM is algebraically equivalent to the ADL, the GECM would—by definition—also be balanced in this example.)

The above discussion suggests that we can use an ADL to estimate the relationship between an integrated Y and stationary X. To test these expectations, we conduct a series of Monte Carlo experiments. We generate an integrated Y with the following DGP:

$$Y_t = Y_{t-1} + \epsilon_{yt},\ \epsilon_{yt}\, N(0,1). \tag{6}$$

We generate the stationary time series X, with the following DGP, where θ equals 0.0 or 0.5:

$$X_t = \theta_x X_{t-1} + \epsilon_{xt},\ \epsilon_{xt}\, N(0,1). \tag{7}$$

Notice that X and Y are independent series. Particularly with dependent series that contain a unit root (as is the case here), the dominant concern in time series literature is the potential for estimating spurious relationships (Granger and Newbold 1974, Grant and Lebo 2016, Yule 1926). Thus, our first simulations seek to identify the percentage of analyses that would *incorrectly* reject the null hypothesis of no relationship between a stationary X and integrated Y with an ADL. As noted above, in light of the recommendations in the *PA* symposium to never mix orders of integration, this approach seems highly problematic. However, if the equation is balanced as we suggest, the false rejection rate in our simulations should only be about 5 percent.

In the following simulations, T is set to 50 and then 1,000. These values allow us to evaluate both a short time series that political scientists often encounter and a long time series that will approximate the asymptotic properties of the series. We use the DGP from Equations 6 and 7, above, to generate 1,000 simulated data sets. Recall that in our stationary series, θ_x equals 0.5 or 0.0 and Y and X are never related. To evaluate the relationship between X and Y, we estimate an ADL model in Equation 4.[13]

Table 1 reports the average estimated relationship across all simulations between X and Y ($\hat{\beta}_1$ and $\hat{\beta}_2$ in Equation 4) and the percent of simulations in which these relationships were statistically significant. The mean estimated relationship is close to zero and the Type I error rate is close to 5 percent. With this ADL specification, when Y is integrated and X is stationary, mixing integrated and stationary time series does not increase the risk of spurious regression.[14]

[11] Although fractionally integrated variables may also be of interest to researchers, this example focuses on stationary and integrated processes, which offer a clear illustration of the consequences of mixing orders of integration.

[12] Banerjee et al. (1993) wrote this in the context of a discussion of equation balance among cointegrated variables, but the point applies equally well in this context.

[13] The ADL is mathematically equivalent to the general error correction model (GECM), so the GECM would produce the same results, as long as the parameters are interpreted correctly (see Enns et al. (2016)).

[14] The simulations reported in Table 1 also indicate that the ADL specification addresses the issue of serially correlated residuals, which would not be the case with an unbalanced regression. When θ_x=0 and $T = 50$, a Breusch-Godfrey test rejects the null of *no* serial correlation just 6.5% of the time. When T=1,000, we find evidence of serially correlated residuals in just 4.6% of the simulations. When θ_x=0.5, the corresponding rates are 6.4% (T=50) and 4.6% (T=1,000).

Table 2: The Percent of Spurious Regressions with a Stationary Y and an Integrated X

	$\theta_y = 0$				$\theta_y = 0.5$			
$T =$	$\hat{\beta}_1$	%	$\hat{\beta}_2$	%	$\hat{\beta}_1$	%	$\hat{\beta}_2$	%
50	-.0023	6.0%	.0013	5.5%	-0.0034	6.4%	0.0015	6.0%
1,000	.0011	5.6%	-0.0010	5.3%	0.0011	5.5%	-0.0009	5.2%

Notes: $\hat{\beta}$ represents the mean coefficient estimate across 1,000 simulations. % represents the percent of the simulations for which we *incorrectly* reject the null hypothesis of no relationship between X and Y. θ_y represents the autoregressive parameter in the DGP of Y.

Results in Table 2 show that the same pattern of results emerges when X is integrated and Y is stationary.[15] Most time series analysis in the political and social sciences could be accused of mixing orders of integration. Thus, the recommendations of the *PA* symposium could be interpreted as calling this research into question. We have shown, however, that mixing orders of integration does not automatically imply an unbalanced equation. It also does not automatically lead to spurious results.

The Rise of the Super Rich: Reconsidering Volscho and Kelly (2012)

We think the foregoing discussion and analyses offer compelling evidence that, despite the range of statements about equation balance in the *PA* symposium, mixing orders of integration when using a GECM/ADL does not automatically pose a problem to researchers. Of course, to a large degree the previous sections reiterate and unpack what econometricians have shown mathematically (e.g., Sims, Stock and Watson 1990), and so may come as little surprise to some readers (especially those who have not read the *PA* Symposium). Here, we use an applied example to illustrate the importance of correctly understanding equation balance. We turn to a recent article by Volscho and Kelly (2012) that analyzes the rapid income growth among the super-rich in the United States (US). They estimate a GECM of pre-tax income growth among the top 1% and find evidence that political, policy, and economic variables influence the proportion of income going those at the top. Critically for our purposes, they include stationary and integrated variables on the right-hand side, which Grant and Lebo (2016, 26) actually single out as a case where the "GECM model [is] inappropriate with mixed orders of integration." Grant and Lebo go on to assert that Volscho and Kelly's "data is a mix of data types (stationary and integrated), so any hypothesis tests will be based on unbalanced equations" (supplementary appendix, p.48). Based on the conclusion that mixing orders of integration produces an unbalanced equation, Grant and Lebo employ fractional error correction technology and find that none of the political or policy variables (and only some economic variables) matter for incomes among the top 1%.

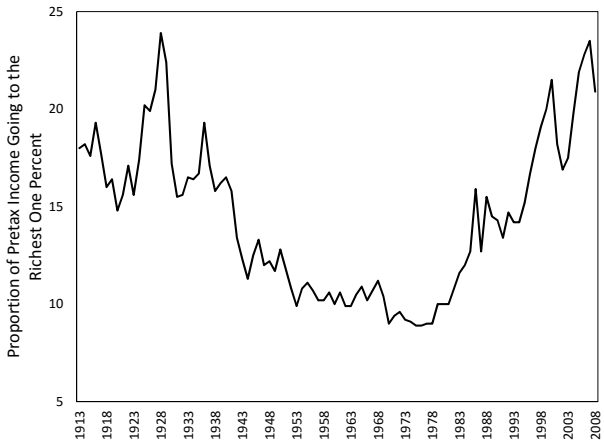


Figure 1: The Top 1 Percent's Share of Pre-tax Income in the United States, 1913 to 2008

These are very different findings, ones with potential policy consequences, and so it is important to reconsider what Volscho and Kelly did—and whether the mixed orders of integration pose a problem for their analysis. To begin our analysis, we present the dependent variable from Volscho and Kelly, the total pre-tax income share of the top 1% for the period between 1913 and 2008.[16] In Figure 1 we can see that income shares start off quite high and then drop and then return to inter-war levels toward the end of the series. The variable thus exhibits none of the trademarks of a

[15] The fact that we do not observe evidence of an increased rate of spurious regression in Table 2, particularly when Y is $AR(1)$, implies that we do not have an equation balance problem. We also find that the simulations in Table 2 tend not to produce serially correlated residuals (we only reject the null of no serial correlation in 6.3% and 5.2% of simulations when T=50 and 4.9% and 3.3% of simulations when T=1,000).

[16] These data, which come from Volscho and Kelly, were originally compiled by Piketty and Saez (2003).

Table 3: Time Series Properties of Variables Analyzed by Volscho and Kelly

Variable	ADF test	Phillips-Perron test	Conclusion
Top 1% Share	0.5121	0.5794	Integrated
Democratic President	0.0462	0.0254	Stationary
Divided Government	0.0082	0.0034	Stationary
Top Marginal Tax Rate	0.1563	0.3351	Integrated
Capital Gains Tax Rate	0.4220	0.5812	Integrated
3-Month Treasury Bill	0.3654	0.2712	Integrated
Trade Openness	0.0608	0.0597	Borderline
Log Real GDP	0.2690	0.2323	Integrated
Real S&P 500 Index	0.2149	0.6427	Integrated
Shiller Home Price Index	0.0000	0.5616	Unclear
Union Membership	0.6710	0.8185	Integrated
% Congressional Dem.	0.0209	0.0147	Stationary

Notes: The Null hypothesis for ADF and Phillips-Perron tests is a unit root. When statistically significant, a trend and/or additional lags were included in the tests.

stationary series, i.e., it is not mean-reverting, and looks to contain a unit root instead. Notice that the same is true for the shorter period encompassed by Volscho and Kelly's analysis, 1949–2008. Augmented Dickey-Fuller (ADF) and Phillips–Perron unit root tests confirm these suspicions, and are summarized in the first row of Table 3.[17]

What about the independent variables? Here, we find a mix (see Table 3). Some variables clearly are nonstationary and also appear to contain unit roots: the capital gains tax rate, union membership, the Treasury Bill rate, Gross Domestic Product (logged), and the Standard and Poor 500 composite index. The top marginal tax rate also is clearly nonstationary and we cannot reject a unit root even when taking into account the secular (trending) decline over time. The results for the Shiller Home Price Index are mixed and trade openness is on the statistical cusp, and there is reason—based on the size of the autoregressive parameter (-0.29) and the fact that we reject the unit root over a longer stretch of time—to assume that the variable is stationary. For the other variables included in the analysis, we reject the null hypothesis of a unit root: Democratic president, and the Percentage of Democrats in Congress. These findings seem to comport with what Volscho and Kelly found (see their supplementary materials).[18]

Volscho and Kelly proceed to estimate a GECM of the top 1% income share including current first differences and lagged levels of the stationary and integrated variables. So far, the diagnostics support their decision (integrated DV, some IVs are integrated, and we find evidence of cointegration).[19] The fact that stationary variables are also included in the model should not affect equation balance. However, in order to evaluate the robustness of Volscho and Kelly's results, we re-consider their data with Pesaran and Shin's ARDL (Autoregressive Distributed Lag) critical bounds testing approach (Pesaran, Shin and Smith 2001). Although political scientists typically refer to the autoregressive distributed lag model as an ADL, Pesaran, Shin and Smith (2001) prefer ARDL. For their bounds test of cointegration, they estimate the model as a GECM.[20]

The ARDL approach is one of the approaches recommended by Grant and Lebo and is especially advantageous in the current context because two critical values are provided, one which assumes all stationary regressors and one which assumes all integrated regressors. Values in between these "bounds" correspond to a mix of integrated and stationary regressors, meaning the bounds approach is especially appropriate when the analysis includes both types of regressors. Grant and Lebo (2016, 19) correctly acknowledge that "With the bounds testing approach, the regressors can be of mixed orders of integration—stationary, non-

[17]These results are consistent with the unit root tests Volscho and Kelly report in the supplementary materials to their article. Grant and Lebo's analysis also supports this conclusion. In their supplementary appendix, Grant and Lebo estimate the order of integration d=0.93 with a standard error of (0.10), indicating they cannot reject the null hypothesis that d=1.0.

[18]Although the dependent variable is pre-tax income, Volscho and Kelly identify several mechanisms that could lead tax rates to influence pre-tax income share (also see, Mertens 2015, Piketty, Saez and Stantcheva 2014). Based on existing research, it also would not be surprising if we observed evidence of a relationship between the top 1 percent's income share and union strength (for recent examples, see Jacobs and Myers 2014, Pontusson 2013, Western and Rosenfeld 2011) and the partisan composition of government (Bartels 2008, Hibbs 1977, Kelly 2009).

[19]When using the error correction parameter in the GECM to evaluate cointegration, the correct Ericsson and MacKinnon (2002) critical values must be used. When doing so, we find evidence of cointegration for Volscho and Kelly's (2012) preferred specification (Model 5).

[20]Recall that the ADL, ARDL, and GECM all refer to equivalent models.

stationary, or fractionally integrated—and the use of bounds allow the researcher to make inferences even when the integration of the regressors is unknown or uncertain."[21] Since Table 3 indicates we have a mix of stationary and integrated regressors, if our critical value exceeds the highest bound, we will have evidence of cointegration.

The ARDL approach proceeds in several steps.[22] First, if the dependent variable is integrated, the ARDL model (which is equivalent to the GECM) is estimated. Next, if the residuals from this model are stationary, an F-test is conducted to evaluate the null hypothesis that the combined effect of all lagged variables in the model equals zero. This F statistic is compared to the appropriate critical values (Pesaran, Shin and Smith 2001). We rely on the small-sample critical values from Narayan (2005). If there is evidence of cointegration, both long and short-run relationships from the initial ARDL (i.e., ADL/GECM) model can be evaluated.

Our analysis focuses on Column 5 from Volscho and Kelly's Table 1, which is their preferred model. The first column of our Table 4, below, shows that we successfully replicate their results. The ARDL analysis appears in Column 2.[23] The key difference between this specification and that of Volscho and Kelly's is that they (based on a Breusch-Godfrey test) employed the Prais-Winsten estimator to correct for serially correlated errors and we do not. Our decision reflects the fact that other tests do not reject the null of white noise, e.g., the Portmanteau (Q) test produces a p-value of 0.12, and it allows us to compare the results with and without the correction. Also note that an expanded model including lagged differenced dependent and independent variables (see Appendix Table 5) produces very similar estimates to those shown in column 2 of Table 4, and a Breusch-Godfrey test indicates that the resulting residuals are uncorrelated.

To begin with, we need to test for cointegration. For this, we compare the F-statistic from the lagged variables (6.54) with the Narayan (2005) upper ($I(1)$) critical value (3.82), which provides evidence of cointegration.[24] The Bounds t-test also supports this inference, as the t-statistic (-7.75) for the Y_{t-1} parameter is greater (in absolute terms) than the $I(1)$ bound tabulated by Pesaran, Shin and Smith (2001, 303). Returning to the results in Column 2, we see that the ARDL approach produces similar conclusions to Column 1. (Philips (2016) uses the ARDL approach to re-consider the first model in Volscho and Kelly's (2012) Table 1 and also obtains similar results.) The coefficients for all but two of the independent variables have similar effects, i.e., the same sign and statistical significance.[25] The exceptions are Divided Government$_{t-1}$ and ΔTrade Openness, for which the coefficients using the two approaches are similar but the standard errors differ substantially. Consistent with the existing research on the subject, we find evidence that economics, politics, and policy matter for the share of income going to the top 1 percent.

Although Grant and Lebo (2016, 18) recommend both the ARDL approach and a three-step fractional error correction model (FECM) approach, they only report the results for the latter in their re-analysis of Volscho and Kelly.[26] It turns out that the two approaches produce very different results. This can be seen in column 3 of Table 4, which reports Grant and Lebo's FECM reanalysis of Volscho and Kelly Model 5 (from Grant and Lebo's supplementary appendix, p.50). With their approach, only the change in stock prices (Real S&P 500 Index) and Trade Openness are statistically significant ($p<.05$) predictors of income shares, though levels of stock prices and trade openness also matter via the FECM component, which captures disequilibria between those variables and lagged income shares. Despite theoretical and empirical evidence suggesting that the marginal tax rate (Mertens 2015, Piketty, Saez and Stantcheva 2014), union strength (Jacobs and Myers 2014, Pontusson 2013, Western and Rosenfeld 2011) and the partisan composition of government (Bartels 2008, Hibbs 1977, Kelly 2009) can influence the pre-tax income of the upper 1 percent, we would conclude that only trade openness and stock prices influence the pre-tax income share of richest Americans. Of course, analysts might reasonably prefer alternative models to the ones Volscho and Kelly estimate, perhaps opting for a more parsimonious specification, allowing endogenous relationships, and/or including alternate lag specifications. The key point is that, given the particular model, the ARDL and three–step FECM produce very different estimates.

[21] It is not clear why Grant and Lebo seemingly contradict their statement that "Mixing together series of various orders of integration will mean a model is misspecified" (p.4) in this context, especially since the ARDL is equivalent to the GECM, but they are correct to do so.

[22] For a concise overview of the ARDL approach, see http://davegiles.blogspot.ca/2013/06/ardl-models-part-ii-bounds-tests.html.

[23] We exactly follow their lag structure and the assumption of a single endogenous variable, which seemingly is incorrect but possibly intractable.

[24] The 5 percent critical value when T=60 with an unrestricted intercept and no trend is 3.823. Narayan (2005) only reports critical values for up to 7 regressors. However, the size of the critical value *decreases* as the number of regressors increases (Narayan 2005, Pesaran, Shin and Smith 2001), so our reliance on the the critical value based on 7 regressors is actually a conservative test of cointegration. We also tested for integration allowing for short-run effects of all integrated variables and we again find evidence of cointegration ($F = 4.32$).

[25] This reveals that explicitly taking into account serial correlation, which Volscho and Kelly did, has modest consequences.

[26] As Grant and Lebo (2016, 18) explain, the three–Step FECM proceeds as follows. First, Y is regressed on X and the residuals are obtained. The fractional difference parameter, d, is then estimated for each of the three series (Y, X, and the residuals). Grant and Lebo explain that if d for the residuals is less than d for both X and Y, then error correction is occurring. If this is the case, the researcher then fractionally differences Y, X, and the residual by each one'ĂŹs own d value. Finally, the researcher regresses the fractionally differenced Y and the fractionally differenced X, and the lag of the fractionally differenced residual (Grant and Lebo 2016, 18). This regression produces the results reported in Column 3 of Table 4.

Table 4: Replication of Volscho & Kelly (2012) Table 1, Column 5 with the ARDL Bounds Test and the Three-Step FECM employed by Grant and Lebo

	(1) V&K Replication	(2) ARDL	(3) G&L 3-Step FECM	
Top 1% Share$_{t-1}$	-0.65*	-0.93*		
	(0.10)	(0.12)		
% Congressional Dem.$_{t-1}$	-0.05*	-0.06*	Δ^d % Congressional Dem.	-0.01
	(0.02)	(0.02)		(0.04)
Divided Government$_{t-1}$	-0.37*	-0.42	Δ^d Divided Government	0.15
	(0.17)	(.24)		(0.36)
Union Membership$_{t-1}$	-0.28*	-0.41*	Δ^d Union Membership	-0.06
	(0.07)	(0.09)		(0.26)
Top Marginal Tax Rate$_{t-1}$	-0.03*	-0.05*	Δ^d Top Marginal Tax Rate	-0.02
	(0.01)	(0.02)		(0.03)
Δ Capital Gains Tax Rate	-0.03	-0.05	Δ^d Capital Gains Tax Rate	-0.07
	(0.03)	(0.04)		(0.05)
Capital Gains Tax Rate$_{t-1}$	-0.06*	-0.08*		
	(0.02)	(0.02)		
3-Month Treasury Bill$_{t-1}$	0.01	0.02	Δ^d 3-Month Treasury Bill	-0.13
	(0.04)	(0.06)		(0.14)
Δ Trade Openness$_t$	0.20*	0.22	Δ^d Trade Openness	0.38*
	(0.01)	(0.12)		(0.19)
Log Real GDP$_{t-1}$	-5.04*	-8.12*	Δ^d Log Real GDP	3.35
	(1.45)	(1.87)		(6.89)
Δ Real S&P 500 Index$_t$	0.06*	0.06*	Δ^d Real S&P 500 Index	0.07*
	(0.01)	(0.01)		(0.01)
Real S&P 500 Index$_{t-1}$	0.03*	0.05*		
	(0.01)	(0.01)		
Shiller Home Price Index$_{t-1}$	0.28*	0.43*	Δ^d Shiller Home Price Index	0.32
	(0.07)	(0.10)		(0.27)
			FECM	-0.35*
				(0.14)
Constant	58.99*	91.13*		0.01
	(14.27)	(18.10)		(0.18)
Adj. R^2	0.76	0.67		0.36

Notes: Regression coefficients with standard errors in parentheses, * $p < .05$. The V&K replication (1) is based on Column 5 of their Table 1 and relies on the Prais-Winsten (GLS) estimator. The G&L results (3) come from the Supplementary Materials (p.50) to Grant and Lebo (2016). FECM reflects the long-term equilibrium relationship of both Trade Openness and Real S&P Composite Index and the other coefficients reflect estimated short run effects based on G&L's Three-Step FECM.

Conclusions

In his contribution to the *PA* symposium, John Freeman wrote, "It now is clear that equation balance is not understood by political scientists" (Freeman 2016, 50). Our goal has been to help clarify misconceptions about equation balance. In particular, we have shown that mixing orders of integration in a GECM/ADL model does *not* automatically lead to an unbalanced equation. As the title of Lebo and Grant's second contribution to the symposium ("Equation Balance and Dynamic Political Modeling") illustrates, equation balance was a central theme of the symposium. Although others have responded to particular criticisms within the *PA* symposium (e.g., Enns et al. (2016)), this article is the first to address the symposium's discussion and recom-

mendations related to equation balance.

Because they are related, it is easy to (erroneously) conclude that mixing orders of integration is synonymous with an unbalanced equation. It would be wrong, however, to reach this conclusion. We have focused on two types of time series: stationary and unit–root series and we have found that situations exist when it is unproblematic—and inconsequential—to mix these types of series (because the equation is balanced).[27]

These results help clarify existing time series research (e.g., Banerjee et al. 1993, Sims, Stock and Watson 1990) by showing that when we use a GECM/ADL to model dynamic processes, even mixed orders of integration can produce balanced equations. The findings also lead to three recommendations for researchers. First, scholars should not automatically dismiss existing time series research that mixes orders of integration. Even when series are of different orders of integration or when the equation transforms variables in a way that leads to different orders of integration, the equation may still be balanced and the model correctly specified. In fact, we identified, and our simulations confirmed, specific scenarios when integrated and stationary time series can be analyzed together. Second, as we showed with our simulations and with our applied example, researchers must evaluate whether they have equation balance based on *both* the univariate properties of their variables *and* the model they specify. Third and finally, our results show that researchers do *not* always need to pre-whiten their data to ensure equation balance. Although pre-whitening time series will sometimes be appropriate, we have shown that this step is not a necessary condition for equation balance. This is important because such data transformations are potentially quite costly, specifically, in the presence of equilibrium relationships. As we saw above, Grant and Lebo's decision to pre-whiten Volscho and Kelly's data with their three-step FECM may be one such example.

Acknowledgments

A previous version of this paper was presented at the Texas Methods Conference, 2017. We would like to thank Neal Beck, Patrick Brandt, Harold Clarke, Justin Esarey, John Freeman, Nate Kelly, Jamie Monogan, Mark Pickup, Pablo Pinto, Randy Stevenson, Thomas Volscho, and two anonymous reviewers for helpful comments and suggestions. All replication materials are available on The Political Methodologist Dataverse site (https://dataverse.harvard.edu/dataverse/tpmnewsletter).

References

Banerjee, Anindya, Juan Dolado, John W. Galbraith and David F. Hendry. 1993. *Co-Integration, Error Correction, and the Econometric Analysis of Non-Stationary Data.* Oxford: Oxford University Press.

Bartels, Larry M. 2008. *Unequal Democracy.* Princeton: Princeton University Press.

Box-Steffensmeier, Janet and Agnar Freyr Helgason. 2016. "Introduction to Symposium on Time Series Error Correction Methods in Political Science." *Political Analysis* 24(1):1–2.

Enns, Peter K., Nathan J. Kelly, Takaaki Masaki and Patrick C. Wohlfarth. 2016. "Don't Jettison the General Error Correction Model Just Yet: A Practical Guide to Avoiding Spurious Regression with the GECM." *Research and Politics* 3(2):1–13.

Ericsson, Neil R. and James G. MacKinnon. 2002. "Distributions of Error Correction Tests for Cointegration." *Econometrics Journal* 5(2):285–318.

Esarey, Justin. 2016. "Fractionally Integrated Data and the Autodistributed Lag Model: Results from a Simulation Study." *Political Analysis* 24(1):42–49.

Freeman, John R. 2016. "Progress in the Study of Nonstationary Political Time Series: A Comment." *Political Analysis* 24(1):50–58.

Granger, Clive W.J., Namwon Hyung and Yongil Jeon. 2001. "Spurious Regressions with Stationary Series." *Applied Economics* 33:899–904.

Granger, Clive W.J. and Paul Newbold. 1974. "Spurious Regressions in Econometrics." *Journal of Econometrics* 26:1045–1066.

Grant, Taylor and Matthew J. Lebo. 2016. "Error Correction Methods with Political Time Series." *Political Analysis* 24(1):3–30.

Hibbs, Jr., Douglas A. 1977. "Political Parties and Macroeconomic Policy." *American Political Science Review* 71(4):1467–1487.

Jacobs, David and Lindsey Myers. 2014. "Union Strength, Neoliberalism, and Inequality." *American Sociological Re-*

[27] Of course, other statistical assumptions must also be satisfied. In other work, we have also considered combined time series that contain both stationary and unit–root properties (Wlezien 2000). We find that when we analyze combined time series with mixed orders of integration, we are able to detect true relationships in the data. These results further highlight the fact that mixed orders of integration do not automatically imply an unbalanced regression.

view 79(4):752–774.

Keele, Luke, Suzanna Linn and Clayton McLaughlin Webb. 2016*a*. "Concluding Comments." *Political Analysis* 24(1):83–86.

Keele, Luke, Suzanna Linn and Clayton McLaughlin Webb. 2016*b*. "Treating Time with All Due Seriousness." *Political Analysis* 24(1):31–41.

Kelly, Nathan J. 2009. *The Politics of Income Inequality in the United States.* New York: Cambridge University Press.

Lebo, Matthew J. and Taylor Grant. 2016. "Equation Balance and Dynamic Political Modeling." *Political Analysis* 24(1):69–82.

Maddala, G.S. and In-Moo Kim. 1998. *Unit Roots, Cointegration, and Structural Change.* 1st ed. New York: Cambridge University Press.

Mankiw, N. Gregory and Matthew D. Shapiro. 1986. "Do We Reject Too Often? Small Sample Properties of Tests of Rational Expectations Models." *Economics Letters* 20(2):139–145.

Mertens, Karel. 2015. "Marginal Tax Rates and Income: New Time Series Evidence." https://mertens.economics.cornell.edu/papers/MTRI_september2015.pdf.

Murray, Michael P. 1994. "A Drunk and Her Dog: An Illustration of Cointegration and Error Correction." *The American Statistician* 48(1):37–39.

Narayan, Paresh Kumar. 2005. "The Saving and Investment Nexus for China: Evidence from Cointegration Tests." *Applied Economics* 37(17):1979–1990.

Pagan, A.R. and M.R. Wickens. 1989. "A Survey of Some Recent Econometric Methods." *The Economic Journal* 99(398):962–1025.

Pesaran, Hashem M., Yongcheol Shin and Richard J. Smith. 2001. "Bounds Testing Approaches to the Analysis of Level Relationships." *Journal of Applied Econometrics* 16(3):289–326.

Philips, Andrew Q. 2016. "Have Your Cake and Eat it Too? Cointegration and Dynamic Inference from Autoregressive Distributed Lag Models." *Working Paper.*

Piketty, Thomas and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118(1):1–39.

Piketty, Thomas, Emmanuel Saez and Stefanie Stantcheva. 2014. "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities." *American Economic Journal: Economic Policy* 6(1):230–271.

Pontusson, Jonas. 2013. "Unionization, Inequality and Redistribution." *British Journal of Industrial Relations* 51(4):797–825.

Sánchez Urribarrí, Raúl A., Susanne Schorpp, Kirk A. Randazzo and Donald R. Songer. 2011. "Explaining Changes to Rights Litigation: Testing a Multivariate Model in a Comparative Framework." *Journal of Politics* 73(2):391–405.

Sims, Christopher A., James H. Stock and Mark W. Watson. 1990. "Inference in Linear Time Series Models with Some Unit Roots." *Econometrica* 58(1):113–144.

Volscho, Thomas W. and Nathan J. Kelly. 2012. "The Rise of the Super-Rich: Power Resources, Taxes, Financial Markets, and the Dynamics of the Top 1 Percent, 1949 to 2008." *American Sociological Review* 77(5):679–699.

Western, Bruce and Jake Rosenfeld. 2011. "Unions, Norms, and the Rise of U.S. Wage Inequality." *American Sociological Review* 76(4):513–537.

Wlezien, Christopher. 2000. "An Essay on 'Combined' Time Series Processes." *Electoral Studies* 19(1):77–93.

Yule, G. Udny. 1926. "Why do we Sometimes get Nonsense-Correlations between Time-Series?-A Study in Sampling and the Nature of Time-Series." *Journal of the Royal Statistical Society* 89:1–63.

Appendix 1 Alternate ARDL Model Specification

Table 5: Alternate Specification of the ARDL Model in Table 4

Top 1% Share$_{t-1}$	-0.88*
	(0.13)
Δ Top 1% Share$_{t-1}$	-0.24*
	(0.10)
% Congressional Dem.$_{t-1}$	-0.06*
	(0.02)
Divided Government$_{t-1}$	-0.42
	(0.22)
Union Membership$_{t-1}$	-0.36*
	(0.09)
Top Marginal Tax Rate$_{t-1}$	-0.04*
	(0.02)
Δ Capital Gains Tax Rate$_t$	-0.08*
	(0.04)
Δ Capital Gains Tax Rate$_{t-1}$	0.06
	(0.04)
Capital Gains Tax Rate$_{t-1}$	-0.11*
	(0.02)
3-Month Treasury Bill$_{t-1}$	-0.01
	(0.06)
Δ Trade Openness$_t$	0.17
	(0.11)
Δ Trade Openness$_{t-1}$	0.14
	(0.11)
Log Real GDP$_{t-1}$	-6.54*
	(1.85)
Δ Real S&P 500 Index$_t$	0.06*
	(0.01)
Δ Real S&P 500 Index$_{t-1}$	0.03*
	(0.01)
Real S&P 500 Index$_{t-1}$	0.04*
	(0.01)
Shiller Home Price Index$_{t-1}$	0.38*
	(0.10)
Constant	77.52*
	(18.17)
Breusch-Godfrey	0.13

Notes: The null hypothesis for the Breusch-Godfrey LM test is *no* autocorrelation.

Lowering the Threshold of Statistical Significance to $p < 0.005$ to Encourage Enriched Theories of Politics

Justin Esarey
Rice University
justin@justinesarey.com

Introduction

A large and interdisciplinary group of researchers recently proposed redefining the conventional threshold of statistical significance from $p < 0.05$ to $p < 0.005$, both two-tailed (Benjamin et al., 2017). The purpose of the reform is to "immediately improve the reproducibility of scientific research in many fields" (p.5); this comes in the context of recent large-scale replication efforts that have uncovered startlingly low rates of replicability among published results (e.g., Klein et al., 2014; Open Science Collaboration, 2015). Recent work suggests that results that meet a more stringent standard for statistical significance will indeed be more reproducible (Johnson, 2013; Esarey and Wu, 2016; Johnson et al., 2017). Reproducibility should be improved by this reform because the *false discovery rate*, or the proportion of statistically significant findings that are null relationships (Benjamini and Hochberg, 1995), is reduced by its implementation. Benjamin et al. (2017) explicitly disavow a requirement that results meet the stricter standard in order to be publishable, but in the past statistical significance has been used as a necessary condition for publication (Sterling, 1959; Sterling, Rosenbaum and Winkam, 1995) and we must therefore anticipate that a redefinition of the threshold for significance may lead to a redefinition of standards for publishability.

As a method of screening empirical results for publishability, the conventional $p < 0.05$ null hypothesis significance test (NHST) procedure lies on a kind of Pareto frontier: it is difficult to improve its qualities on one dimension (i.e., increasing the replicability of published research) without degrading its qualities on some other important dimension. This makes it hard for any proposal to displace the conventional NHST: such a proposal will almost certainly be worse in some ways, even if it is better in others. For moving the threshold for statistical significance from 0.05 to 0.005, the most obvious tradeoff is that the *power* of the test to detect non-null relationships is harmed at the same time that the *size* of the test (i.e., its propensity to mistakenly reject a true null hypothesis) is reduced. The usual way of increasing the power of a study is to increase the sample size, N; given the fixed nature of historical time-series cross-sectional data sets and the limited budgets of those who use experimental methods, this means that many researchers may feel forced to accept less powerful studies and therefore have fewer opportunities to publish. Additionally, reducing the threshold for statistical significance can exacerbate *publication bias*, i.e., the extent to which the published literature exaggerates the magnitude of an relationship by publishing only the largest findings from the sampling distribution of a relationship (Sterling, Rosenbaum and Winkam, 1995; Scargle, 2000; Schooler, 2011; Esarey and Wu, 2016).

Simply accepting lower power in exchange for a lower false discovery rate would increase the burden on the most vulnerable members of our community: assistant professors and graduate students, who must publish enough work in a short time frame to stay in the field. However, there is a way of maintaining adequate power using $p < 0.005$ significance tests without dramatically increasing sample sizes: design studies that conduct conjoint tests of multiple predictions from a single theory (instead of individual tests of single predictions). When K-many statistically independent tests are performed on pre-specified hypotheses that must be jointly confirmed in order to support a theory, the chance of simultaneously rejecting them all by chance is α^K where $p < \alpha$ is the critical condition for statistical significance in an individual test. As K increases, the α value for each individual study can fall and the overall power of the study often (though not always) increases. It is important that hypotheses be specified *before* data analysis and that failed predictions are reported; simply conducting many tests and reporting the statistically significant results creates a multiple comparison problem that inflates the false positive rate (Sidak, 1967; Abdi, 2007). Because it is usually more feasible to collect a greater depth of information about a fixed-size sample rather than to greatly expand the sample size, this version of the reform imposes fewer hardships on scientists at the beginning of their careers.

I do not think that an NHST with a lowered α is the best choice for adjudicating whether a result is statistically meaningful; approaches rooted in statistical decision theory and with explicit assessment of out-of-sample prediction have advantages that I consider decisive. However, if reducing α prompts us to design research stressing the simultaneous testing of multiple theoretical hypotheses, I believe this change would improve upon the status quo—even if passing this more selective significance test becomes a requirement for publication.

False discovery rates and the null hypothesis significance test

Based on the extant literature, I believe that requiring published results to pass a two-tailed NHST with $\alpha = 0.05$ is at least partly responsible for the "replication crisis" now underway in the social and medical sciences (Wasserstein and Lazar, 2016). I also anticipate that studies that can meet a lowered threshold for statistical significance will be

more reproducible. In fact, I argued these two points in a recent paper (Esarey and Wu, 2016). Figure 1 reproduces a relevant figure from that article; the y-axis in that figure is the *false discovery rate* (Benjamini and Hochberg, 1995), or FDR, of a conventional NHST procedure with an α given by the value on the x-axis. Written very generically, the false discovery rate is:

$$\begin{aligned} FDR &= \Pr(\text{null hypothesis is true} \mid \text{result is stat. sig.}) \\ &= \frac{A}{A+B} \\ A &= \Pr(\text{result is stat. sig.} \mid \text{null hypothesis is true}) \\ &\quad \times \Pr(\text{null hypothesis is true}) \\ B &= \Pr(\text{result is stat. sig.} \mid \text{null hypothesis is false}) \\ &\quad \times (1 - \Pr(\text{null hypothesis is true})) \end{aligned}$$

or, the proportion of statistically significant results ("discoveries") that correspond to true null hypotheses. A is a function of the *power* of the test, Pr(stat. sig.|null hypothesis is true). B is a function of the *size* of the test, Pr(stat. sig.|null hypothesis is false). Both A and B are a function of the underlying proportion of null hypotheses being proposed by researchers, Pr(null hypothesis is true).

As Figure 1 shows, when Pr(null hypothesis is true) is large, there is a very disappointing FDR among results that pass a two-tailed NHST with $\alpha = 0.05$. For example, when 90% of the hypotheses tested by researchers are false leads (i.e., Pr(null hypothesis is true) = 0.9), we may expect nearly $\approx 30\%$ of discoveries to be false when all studies have perfect power (i.e., Pr(stat. sig.|null hypothesis is false) = 1). Two different studies applying disparate methods to data from the Open Science Collaboration's replication project data (2015) have determined that approximately 90% of researcher hypotheses are false (Johnson et al., 2017; Esarey and Liu, 2017); consequently, an FDR in the published literature of at least 30% is attributable to this mechanism. Even higher FDRs will result if studies have less than perfect power.

By contrast, Figure 1 *also* shows that setting $\alpha \approx 0.005$ would greatly reduce the false discovery rate, even when the proportion of null hypotheses posed by researchers is extremely high. For example, if 90% of hypotheses proposed by researchers are false, the lower bound FDR in the published literature among studies meeting the higher standard would be about 5%. Although non-null results are not always perfectly replicable in underpowered studies (and null results have a small chance of being successfully replicated), a reduction in the FDR from $\approx$ 30% to $\approx$ 5% would almost certainty and drastically improve the replicability of published results.

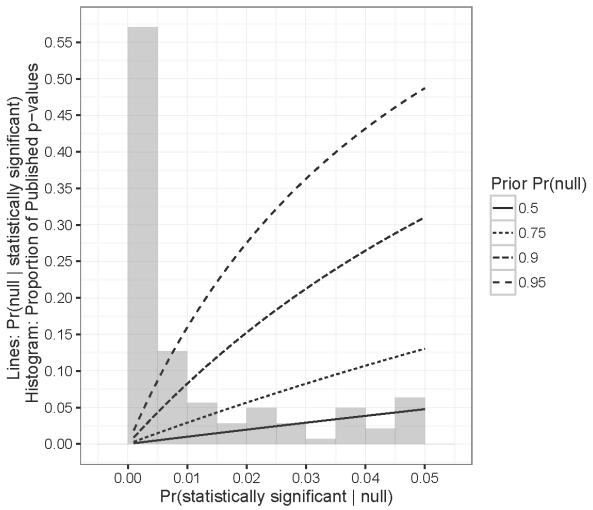


Figure 1: Expected lower bound FDR calculations as a function of the threshold for statistical significance and the background (prior) probability that a proposed hypothesis is false (Figure 2 from Esarey and Wu (2016)). The false discovery rate, Pr(null hypothesis is true|result is stat. sig.), is on the y-axis; the x-axis is the p-value threshold for statistical significance, α. The histogram shows the proportion of p-values in bins of width 0.005 for 142 published and statistically significant results in a data set of articles from the *American Political Science Review* vols. 102-107 and *American Journal of Political Science* vols. 54-57. All studies are assumed to have Pr(stat. sig.|null hypothesis is false) = 1 in order to achieve a lower bound for the FDR estimate.

The Pareto frontier of adjudicating statistically meaningful results

Despite this and other known weaknesses of the current NHST, it lies on or near a Pareto frontier of possible ways to classify results as "statistically meaningful" (one aspect of being suitable for publication). That is, it is challenging to propose a revision to the NHST that will bring improvements without also bringing new or worsened problems, some of which may disproportionately impact certain segments of the discipline. Consider some dimensions on which we might rate a statistical test procedure, the first few of which I have already discussed above:

1. the *false discovery rate* created by the procedure, itself a function of:
   (a) the *power* of the procedure to detect non-zero relationships, and
   (b) the *size* of the procedure, its probability of rejecting true null hypotheses;

2. the degree of *publication bias* created by the procedure, i.e., the extent to which the published literature exaggerates the size of an effect by publishing only the largest findings from the sampling distribution of a relationship (Sterling, Rosenbaum and Winkam, 1995; Scargle, 2000; Schooler, 2011; Esarey and Wu, 2016);

3. the *number of researcher assumptions* needed to execute the procedure, a criterion related to the *consistency* of the standards implied by use of the test; and

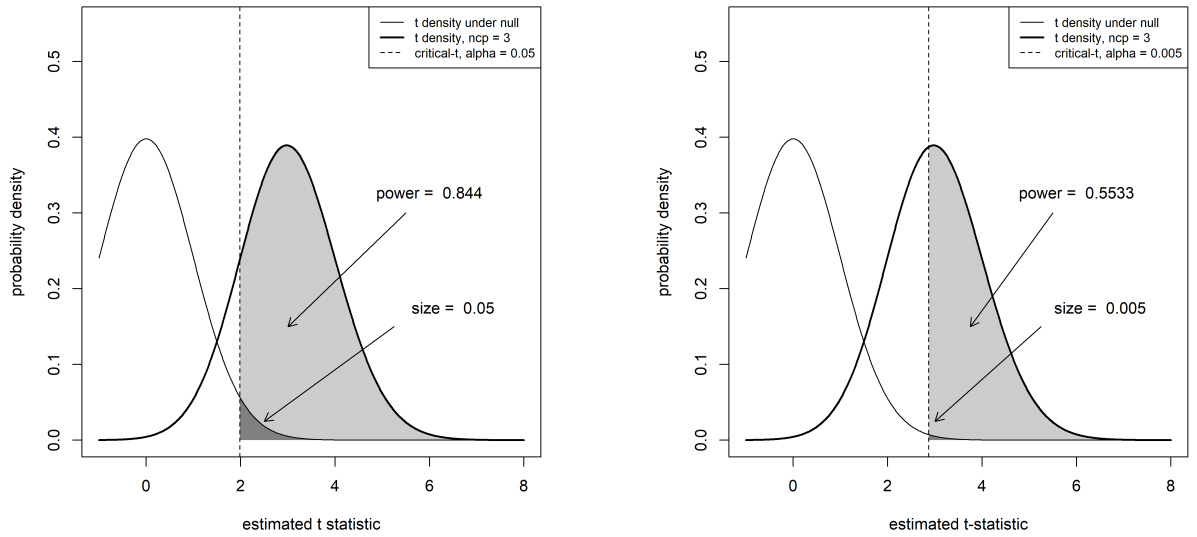
4. the *complexity*, or ease of use and interpretability, of the procedure.

It is hard to improve on the NHST in one of these dimensions without hurting performance in another area. In particular, lowering α from 0.05 to 0.005 would certainly lower the power of most researchers' studies and might increase the publication bias in the literature.

Size/power tradeoffs of lowering α

Changing the α of the NHST from 0.05 to 0.005 is a textbook example of moving along a Pareto frontier because the size and the power of the test are in direct tension with one another. Because powerful research designs are more likely to produce publishable results when the null is false, maintaining adequate study power is especially critical to junior researchers who need publications in order to stay in the field and advance their careers. The size/power tradeoff is depicted in Figure 2.

Figure 2 depicts two size and power analyses for a coefficient of interest β. I assume that $\sigma_{\hat{\beta}}$, the standard deviation of the sampling distribution of β, is equal to 1; I also assume 100 degrees of freedom (typically corresponding to a sample size slightly larger than 100. In the left panel (Figure 2a), the conventional NHST with $\alpha = 0.05$ is depicted. In the right panel (Figure 2b), the Benjamin et al. (2017) proposal to decrease α to 0.005 is shown. The power analyses assume that the true $\beta = 3$, while the size analyses assume that $\beta = 0$.



(a) Conventional NHST, two-tailed $\alpha = 0.05$

(b) Proposed NHST, two-tailed $\alpha = 0.005$

Figure 2: Tradeoff in size and power when α is reduced from 0.05 to 0.005. Size and power calculations for a coefficient of interest β include lower tail areas not depicted in figure. All calculations assume 100 degrees of freedom. The size calculations assume $\beta = 0$ and $\sigma_\beta = 1$ (i.e., a standard t-density). The power calculations assume $\beta = 3$ and $\sigma_{\hat{\beta}} = 1$ (i.e., a t-density with non-centrality parameter $\lambda = 3$).

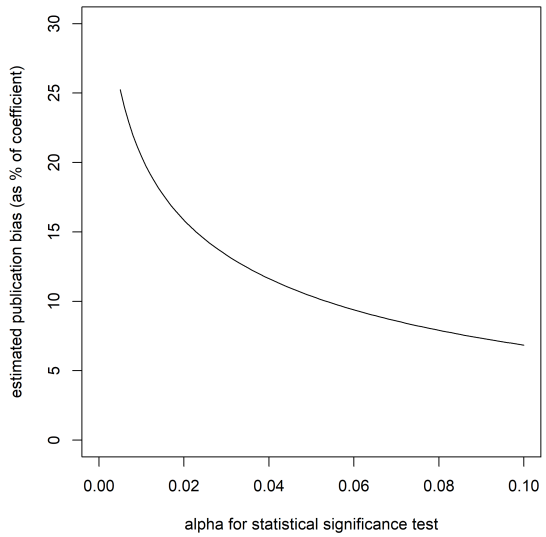
The most darkly shaded area under the right hand tail of the t-distribution under the null is the probability of incorrectly rejecting a true null hypothesis, α. As the figure shows, it is impossible to shrink α without simultaneously shrinking the lighter shaded area, where the lighter shading depicts the power of a hypothesis test to correctly reject a false null hypothesis. This tradeoff can be more or less severe (depending on β and $\sigma_{\hat{\beta}}$), but always exists. The false discovery rate will still (almost always) improve when α falls from 0.05 to 0.005, despite the loss in power, but many more true relationships will not be discovered under the latter standard.

Increased publication bias associated with lowered α

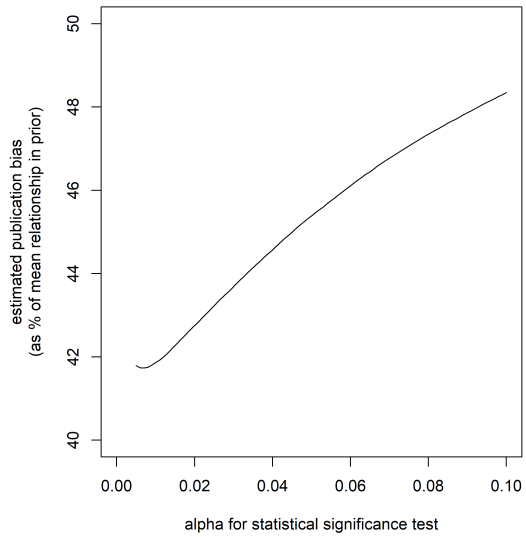
The size/power tradeoff is not the only compromise associated with lowering α: in some situations, decreased α can increase the publication bias associated with using statistical significance tests as a filter for publication. Of course, Benjamin et al. (2017) explicitly disavow using an NHST with lowered α as a necessary condition for publication; they say that "results that would currently be called 'significant' but do not meet the new threshold should instead be called 'suggestive'" (p. 5). However, given the fifty year history of of requiring results to be statistically significant to be publishable (Sterling, 1959; Sterling, Rosenbaum and Winkam, 1995), we must anticipate that this pattern could continue into the future.

Consider Figure 3, which shows two perspectives on how decreased α might impact publication bias. The left panel (Figure 3a) shows the percentage difference in magnitude between a true coefficient $\beta = 3$ and the average statistically significant coefficient using an NHST with various values of α.[1] The figure shows that decreased values of α increase the degree of publication bias; this occurs because stricter significance tests tend to reject only the largest estimates from a sampling distribution, as also shown in Figure 2. On the other hand, the right panel (Figure 3b) shows the percentage difference in magnitude from a true coefficient drawn from a spike-and-normal prior density of coefficients:

$$f(\beta) \sim \mathbb{1}(\beta = 0) * 0.5 + \phi(\mu = 0, \sigma = 3) * 0.5$$



(a) Bias of Single Coefficient, $\beta = 3$



(b) Bias of Distribution of Coefficients, $f(\beta) \sim \mathbb{1}(\beta = 0) * 0.5 + \phi(\mu = 0, \sigma = 3) * 0.5$

Figure 3: Publication bias created by statistical significance tests with varying α. Each graph is created by drawing ten million samples (representing $\hat{\beta}$) from either (a) a t-density with non-centrality parameter $\lambda = 3$ and 100 degrees of freedom, or (b) a draw from a spike-and-normal prior density $f(\beta)$ plus sampling variation (represented by a draw from a standard t-density with 100 degrees of freedom). The y-axis shows the mean value of $(\hat{\beta} - \beta)/\mu_\beta$ for statistically significant values of $\hat{\beta}$ given the α indicated on the x-axis where μ_β is the mean value of the true value of β.

[1] Specifically, I measure $E[(\hat{\beta} - 3)/3]$, where $\hat{\beta}$ is a statistically significant estimate.

where there is a substantial (50%) chance of a null relationship being studied by a researcher; $\mathbb{1}(\beta = 0)$ is the indicator function for a null relationship.[2] In this case, increasingly strict significance tests tend to *decrease* publication bias, because the effect of screening out null relationships (which are greatly overestimated by any false positive result) is stronger than the effect of estimating only a portion of the sampling distribution of non-null relationships (which drives the result in the left panel).

Adapting to a lowered α

If the proposal of Benjamin et al. (2017) is simply to accept lowered power in exchange for lower false discovery rates, it is a difficult proposal to accept. First and foremost, assistant professors and graduate students must publish a lot of high-quality research in a short time frame and may be forced to leave the discipline if they cannot; higher standards that are an inconvenience to tenured faculty may be harmful to them unless standards for hiring and tenure adapt accordingly. In addition, even within a particular level of seniority, this reform may unlevel the playing field. Political scientists in areas that often observational data with essentially fixed N, such as International Relations, would be disproportionately affected by such a change: more historical data cannot be created in order to raise the power of a study. Among experimenters and survey researchers, those with smaller budgets would also be disproportionately affected: they cannot afford to simply buy larger samples to achieve the necessary power. Needless to say, the effect of this reform on the scientific ecosystem would be difficult to predict and not necessarily beneficial; at first glance, such a reform seems to benefit the most senior scholars and people at the wealthiest institutions.

However, I believe that acquiescence to lower power is *not* the only option. It may be difficult or impossible for researchers to collect larger N, but it is considerably easier for them to measure a larger number of variables of interest K from their extant samples. This creates the possibility that researchers can spend more time developing and enriching their theories so that these theories make multiple predictions. Testing these predictions jointly typically allows for much greater power than testing any single prediction alone, as long as any prediction is clearly laid out *prior to analysis* and *all failed predictions are reported* in order to avoid a multiple comparison problem (Sidak, 1967; Abdi, 2007); predictions must be specified in advance and failed predictions must be reported because simply testing numerous hypotheses and reporting any that were confirmed tends to generate an excess of false positive results.

When K-many statistically independent hypothesis tests are performed using a significance threshold of α and *all* must be passed in order to confirm a theory,[3] the chance of simultaneously rejecting them all by chance is $\tilde{\alpha} = \alpha^K$. Fixing $\tilde{\alpha} = 0.005$, it is clear that increasing K allows the individual test's α to be higher.[4] Specifically, α must be equal to $(\tilde{\alpha})^{1/K}$ in order to achieve the desired size. That means that two statistically independent hypothesis tests can have their individual $\alpha \approx 0.07$ in order to achieve a joint size of 0.005; this is a *lower* standard on the individual level than the current 0.05 convention. When $k = 3$, the individual $\alpha \approx 0.17$.

When will conducting a joint test of multiple hypotheses yield greater power than conducting a single test? If two hypothesis tests are statistically independent[5] and conducted as part of a joint study, this will occur when:

$$P_1(\tilde{\alpha}) < P_1(\tilde{\alpha}^{1/2})P_2(\tilde{\alpha}^{1/2}) \tag{1}$$

$$\frac{P_1(\tilde{\alpha})}{P_1(\tilde{\alpha}^{1/2})} < P_2(\tilde{\alpha}^{1/2}) \tag{2}$$

$$P_k(a) = \tau_k(-t^\star(a)) + (1 - \tau_k(t^\star(a))))$$

Here, τ_k is the non-central cumulative t-density corresponding to the sampling distribution of β_k, $k \in \{1, 2\}$; I presume that k is sorted so that tests are in descending order of power. $t^\star(a)$ is the positive critical $t-$statistic needed to create a single two-tailed hypothesis test of size a. The left hand side of equation (1) is the power of the single hypothesis test with the greatest power; the right hand side of (1) is the power of the joint test of two hypotheses. As equation (2) shows, the power of the joint test is larger when the proportional change in power for the test of β_1 is less than the power for the test of β_2. Whether this condition is met depends on many factors, including the magnitude and variability of β_1 and β_2.

To illustrate the potential for power gains, I numerically calculated the power of joint tests with size $\tilde{\alpha} = 0.005$ for the case of one, two, and three statistically independent individual tests. For this calculation, I assumed three relationships $\beta = \beta_k$ with equal magnitude and sign, $k \in \{1, 2, 3\}$ with $\sigma_{\hat{\beta}_k} = 1$; thus, each one of the three tests has identical power when conducted individually. I then calculated the power of each joint test $\left(\tau(-t^\star(\tilde{\alpha})^{1/k}) + (1 - \tau(t^\star((\tilde{\alpha})^{1/k})))\right)^k$ for each value of k and for varying values of β, where τ is the non-central cumulative t-density with non-centrality parameter equal to β. Note that, as before, I define $t^\star(a)$ as the

[2] Here, I measure $E[(\hat{\beta} - \beta)/\mu_{beta}]$, where μ_β is the mean of $f(\beta)$.

[3] The points in this paragraph are similar to those made by Esarey and Sumner (2017, pp. 15-19).

[4] For statistically correlated tests, the degree to which $\alpha > \tilde{\alpha}$ is smaller; at the limit where all the hypothesis tests are perfectly correlated, $\alpha = \tilde{\alpha}$.

[5] Correlated significance tests require the creation of a joint distribution τ on the right hand side of equation (1) and the determination of a critical value $t^\star$ such that $\int_{t^\star}^{\infty}\int_{t^\star}^{\infty} \tau(t_1, t_2)dt_1dt_2 = \tilde{\alpha}$; while practically important, this analysis is not as demonstratively illuminating as the case of statistically independent tests.

critical t-statistic for a single two-tailed test with size a. The result is illustrated in Figure 4.

Figure 4 shows that joint hypothesis testing creates substantial gains in power over single hypothesis testing for most values of β. There is a small amount of power loss near the tails (where $\beta \approx 0$ and $\beta \approx 6$), but this is negligible.

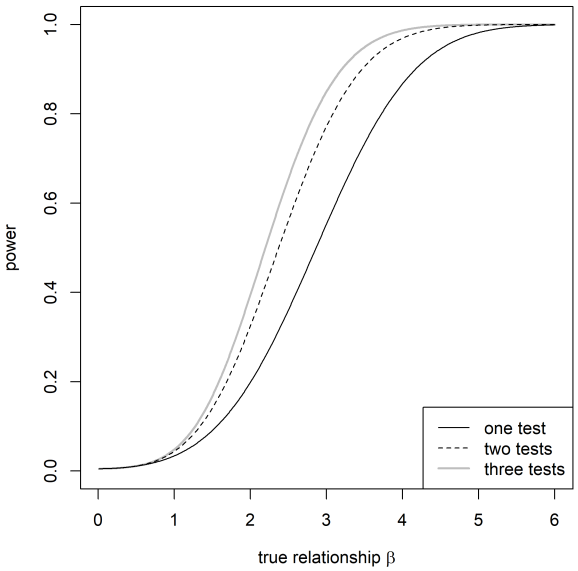


Figure 4: Power calculations for joint tests of one, two, and three statistically independent null hypotheses. The joint test of K-many hypotheses derived from a single theory is deemed passed when *all* relationships are statistically significant in individual two-tailed tests with size $\tilde{\alpha}^{1/K}$, where $\tilde{\alpha}$ is the target size for the joint hypothesis test.

Conclusion: the NHST with lowered α is an improvement, if we enrich our theories

There are reasons to be skeptical of the Benjamin et al. (2017) proposal to move the NHST threshold for statistical significance from $\alpha = 0.05$ to $\alpha = 0.005$. First, there is substantial potential for adverse effects on the scientific ecosystem: the proposal seems to advantage senior scholars at the most prominent institutions in fields that do not rely on fixed-N observational data. Second, the NHST with a reduced α is not my ideal approach to adjudicating which results are statistically meaningful; I believe it is more advantageous for political scientists to adopt a statistical decision theory-oriented approach to inference[6] and give greater emphasis to cross-validation and out-of-sample prediction.

However, based on the evidence presented here and in related work, I believe that moving the threshold for statistical significance from $\alpha = 0.05$ to $\alpha = 0.005$ would benefit political science *if* we adapt to this reform by developing richer, more robust theories that admit multiple predictions. Such a reform would reduce the false discovery rate without reducing power or unduly disadvantaging underfunded scholars or subfields that rely on historical observational data, even if meeting the stricter standard for significance became a necessary condition for publication. It would also force us to focus on improving our body of theory; our extant theories lead us to propose hypotheses that are wrong as much as 90% of the time (Johnson et al., 2017; Esarey and Liu, 2017). Software that automates the calculation of appropriate critical t-statistics for correlated joint hypothesis tests would make it easier for substantive researchers to make this change, and ought to be developed in future work. This software has already been created for joint tests involving interaction terms in generalized linear models by Esarey and Sumner (2017), but the procedure needs to be adapted for the more general case of any type of joint hypothesis test.

Acknowledgments

I thank Jeff Grim, Martin Kavka, Tim Salmon, and Mike Ward for helpful comments on a previous draft of this paper, particularly in regard to the effect of lowered α on junior scholars and the question of whether statistical significance should be required for publication.

References

Abdi, Herve. 2007. The Bonferonni and Sidak Corrections for Multiple Comparisons. In *Encyclopedia of Measurement and Statistics*, ed. Neil Salkind. Thousand Oaks, CA: Sage.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcom Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, James Holland Jones, Daniel J. Hruschka,

[6] In a paper with Nathan Danneman (2015), I show that a simple, standardized approach could reduce the rate of false positives without harming our power to detect true positives (see Figure 4 in that paper). Failing this, I would prefer a statistical significance decision explicitly tied to expected replicability, which requires information about researchers' propensity to test null hypotheses as well as their bias toward positive findings (Esarey and Liu, 2017). These changes would increase the *complexity* and the *number of researcher assumptions* of a statistical assessment procedure relative to the NHST, but not (in my opinion) to a substantial degree.

Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson. 2017. "Redefine Statistical Significance." *Nature Human Behavior* Forthcoming:1–18.

Benjamini, Y. and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Esarey, Justin and Ahra Wu. 2016. "Measuring the Effects of Publication Bias in Political Science." *Research & Politics* 3(3):1–9.

Esarey, Justin and Jane Lawrence Sumner. 2017. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." *Comparative Political Studies* Forthcoming:1–39.

Esarey, Justin and Nathan Danneman. 2015. "A Quantitative Method for Substantive Robustness Assessment." *Political Science Research and Methods* 3(1):95–111.

Esarey, Justin and Vera Liu. 2017. "A Prospective Test for Replicability and a Retrospective Analysis of Theoretical Prediction Strength in the Social Sciences." Poster presented at the 2017 Texas Methods Meeting at the University of Houston. URL: http://jee3.web.rice.edu/replicability-package-poster.pdf accessed 8/1/2017.

Johnson, Valen E. 2013. "Revised standards for statistical evidence." *Proceedings of the National Academy of Sciences* 110(48):19313–19317.

Johnson, Valen E, Richard D. Payne, Tianying Wang, Alex Asher and Soutrik Mandal. 2017. "On the Reproducibility of Psychological science." *Journal of the American Statistical Association* 112(517):1–10.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Stepan Bahnik, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka and Brian A. Nosek. 2014. "Investigating Variation in Replicability." *Social Psychology* 45(3):142–152.

Nelson, Leif D., Joseph P. Simmons and Uri Simonsohn. 2012. "Let's Publish Fewer Papers." *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory* 23(3):291–293.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):aac4716.

Scargle, Jeffrey D. 2000. "Publication Bias: The 'File-Drawer' Problem in Scientific Inference." *Journal of Scientific Exploration* 14:91–106.

Schooler, Jonathan. 2011. "Unpublished Results Hide the Decline Effect." *Nature* 470:437.

Sidak, Zbynek. 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62(318):626–633.

Sterling, T.D., W. L. Rosenbaum and J.J. Winkam. 1995. "Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa." *The American Statistician* 49:108–112.

Sterling, Theodore D. 1959. "Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance–or Vice Versa." *Journal of the American Statistical Association* 54(285):30–34.

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70(2):129–133.

Pitfalls when Estimating Treatment Effects Using Clustered Data

James G. MacKinnon
Queen's University
jgm@econ.queensu.ca

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

Introduction

There is a large and rapidly growing literature on inference with clustered data, that is, data where the disturbances (error terms) are correlated within clusters. The clusters might be associated with, for example, jurisdictions, schools, hospitals, industries, or time periods. Cameron and Miller (2015) provides a very good survey. However, the literature is growing rapidly. More recent papers include Imbens and Kolesár (2016), MacKinnon and Webb (2017*b*), Carter, Schnepel, and Steigerwald (2017), Pustejovsky and Tipton (2017), Djogbenou, MacKinnon, and Nielsen (2017), MacKinnon, Nielsen, and Webb (2017), and Esarey and Menger (2017). Most of these papers are written by and for economists, but the Esarey-Menger paper is specifically aimed at researchers in political science.

Since the theoretical justification for cluster-robust standard errors is asymptotic, it is evident that we need sufficiently large samples if we are to make valid inferences. It has long been recognized that what matters is not the number of observations but rather the number of clusters. Based on the limited simulation evidence available at the time it was written, Angrist and Pischke (2008, Chapter 8) suggested, not entirely seriously, that it is safe to use cluster-robust standard errors whenever there are at least 42 clusters. But the evidence on which they based this suggestion was for the best-case scenario of equal-size clusters and continuous regressors.

More recent work suggests that, by itself, the number of clusters does not tell us whether inference is likely to be reliable. MacKinnon and Webb (2017*b*) shows by simulation that using cluster-robust t statistics can be quite unreliable when there are either 50 or 100 clusters that are proportional to the sizes of U.S. states (with each state appearing twice in the latter case). In general, it seems to be the case that, as cluster sizes become more unequal, inference becomes less reliable.

Perhaps more surprisingly, MacKinnon and Webb (2017*b*) also show that, when the regressor of interest is a treatment dummy, cluster-robust standard errors can be very much too small whenever the number of treated clusters is small. This result is obtained by theory and confirmed by simulation. It holds even when the total number of clusters is very large. Previous simulation evidence on this point may be found in Bell and McCaffrey (2002) and Conley and Taber (2011). Since many applied studies in economics and political science involve either pure treatment models or difference-in-differences (DiD) models, this finding has serious implications for applied work in both fields.

Bootstrap methods typically perform better than t tests, but they can also yield very misleading inferences in some cases. In particular, what would otherwise be the best variant of the wild bootstrap can underreject extremely severely when the number of treated clusters is very small. Other bootstrap methods can overreject extremely severely in that case.

In Section 2, we briefly review the key ideas of cluster-robust covariance matrices and standard errors. In Section 3, we then explain why inference based on these standard errors can fail when there are few treated clusters. In Section 4, we discuss bootstrap methods for cluster-robust inference. In Section 5, we report (graphically) the results of several simulation experiments which illustrate just how severely both conventional and bootstrap methods can overreject or underreject when there are few treated clusters. In Section 6, the implications of these results are illustrated using an empirical example from Burden, Canon, Mayer, and Moynihan (2017). Finally, Section 7 concludes and provides some recommendations for empirical work.

Cluster-Robust Standard Errors

We are concerned with the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}') = \boldsymbol{\Omega}, \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $N \times 1$ vectors of observations and disturbances, $\boldsymbol{X}$ is an $N \times k$ matrix of covariates, $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector, and the $N \times N$ covariance matrix $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Omega}_2 & \dots & \mathbf{O} \\ \vdots & \vdots & & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & \boldsymbol{\Omega}_G \end{bmatrix}. \tag{2}$$

There are G clusters, indexed by g, with N_g observations in the g^{th} cluster. For notational convenience, the observations are assumed to be ordered by cluster, although this is not necessary in practice. The $N_g \times N_g$ matrix $\boldsymbol{\Omega}_g$ is positive definite. It is the covariance matrix for the observations belonging to the g^{th} cluster. Thus $\boldsymbol{\Omega}$ is block-diagonal, with G diagonal blocks that correspond to the G clusters.

The true covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}}$ for the model (1) is

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \end{aligned} \tag{3}$$

where $\boldsymbol{X}_g$ denotes the N_g rows of $\boldsymbol{X}$ that belong to the g^{th} cluster. The most widely-used cluster-robust variance

estimator, or CRVE, is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \frac{G(N-1)}{(G-1)(N-k)}(\boldsymbol{X}'\boldsymbol{X})^{-1} \times \left(\sum_{g=1}^{G} \boldsymbol{X}_g'\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{4}$$

The first factor here is asymptotically negligible, but it always makes the CRVE larger when G and N are finite.[1] Covariance matrix estimators like equation (4) are often referred to as "sandwich estimators" because there are two identical pieces of "bread" on the outside and a "filling" in the middle. The filling in the sandwich on the right-hand side of equation (4) is evidently intended to estimate the corresponding factor in equation (3). In both cases, the filling involves a sum of G $k \times k$ matrices. In the case of (4), each of these matrices has rank one. Therefore, the matrix (4) can have rank at most G.

The CRVE (4) is often called CV_1, because it is the analog of the heteroskedasticity-robust covariance matrix estimator HC_1; see MacKinnon (2012). A more complicated CRVE, often called CV_2, which is the analog of the HC_2 estimator studied in MacKinnon and White (1985), was proposed in Bell and McCaffrey (2002) and has recently been advocated by Imbens and Kolesár (2016); see also Pustejovsky and Tipton (2017). CV_2 generally has better finite-sample properties than CV_1. It does not solve the problems associated with few treated clusters, but it can make them less severe. Unfortunately, CV_2 is considerably more expensive to compute than CV_1 when the clusters are large.[2] Although we do not discuss CV_2 further, it should certainly be considered for samples of moderate size.

The most common way to make inferences about any element of $\boldsymbol{\beta}$, say β_k, is to divide the OLS estimate $\hat{\beta}_k$ by the square root of the k^{th} diagonal element of the CRVE (4) and compare the resulting t statistic to the $t(G-1)$ distribution. This procedure, which can be much more conservative than using the standard normal distribution when G is small, was suggested in Bester, Conley, and Hansen (2011). There are also several (moderately complicated) procedures for calculating the degrees of freedom based on $\boldsymbol{X}$ and the assumed cluster structure; see Bell and McCaffrey (2002), Imbens and Kolesár (2016), Young (2016), and Carter, Schnepel, and Steigerwald (2017). These typically yield non-integer degrees of freedom that are even smaller than $G-1$.

Inference with Few Treated Clusters

The fundamental problem with CRVE-based inference when there are few treated clusters is that, in such cases, the residuals typically provide very poor estimates of the disturbances for the treated clusters. Following MacKinnon and Webb (2017*b*, Section 6), we consider the pure treatment model

$$y_{gi} = \beta_1 + \beta_2 d_{gi} + u_{gi}, \quad i = 1\ldots, N_g, \quad g = 1, \ldots, G, \tag{5}$$

where d_{gi} equals 1 for the first G_1 clusters and 0 for the remaining $G_0 = G - G_1$ clusters. In this model, every observation in the g^{th} cluster is either treated ($d_{gi} = 1$) or not treated ($d_{gi} = 0$). The analysis would be more complicated if we included additional regressors, or allowed only some observations within treated clusters to be treated, but it would not change in any fundamental way.

From expression (4), it is not hard to show that, if we omit the initial scalar factor, the CRVE for $\hat{\beta}_2$, the OLS estimator of β_2 in equation (5), is equal to

$$\frac{\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)}{\big((\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})\big)^2}. \tag{6}$$

Here $\boldsymbol{d}$ denotes the vector with typical element d_{gi}, $\boldsymbol{d}_g$ is the subvector corresponding to cluster g, $\bar{d}$ is the mean of the d_{gi} (that is, the proportion of the observations that are treated), and $\boldsymbol{\iota}$ and $\boldsymbol{\iota}_g$ are vectors of 1s, of lengths N and N_g, respectively. The numerator of (6) is essentially the filling in the CRVE sandwich, and the denominator is the square of the bread. Only the numerator depends on the residuals.

Expression (6) would provide a good estimate of $\mathrm{Var}(\hat{\beta}_2)$ if its numerator provided a good estimate of the filling in the sandwich (3) specialized to the case of β_2 in the treatment model (5). In scalar notation, this numerator can be written as

$$(1-\bar{d})^2 \sum_{g=1}^{G_1}\left(\sum_{i=1}^{N_g}\hat{u}_{gi}\right)^2 + \bar{d}^2 \sum_{g=G_1+1}^{G}\left(\sum_{i=1}^{N_g}\hat{u}_{gi}\right)^2. \tag{7}$$

Expression (7) is supposed to estimate the quantity

$$(1-\bar{d})^2 \sum_{g=1}^{G_1}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\omega_{ij}^g + \bar{d}^2 \sum_{g=G_1+1}^{G}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\omega_{ij}^g, \tag{8}$$

where ω_{ij}^g is the ij^{th} element of $\boldsymbol{\Omega}_g$. But it does a terrible job of doing so when the number of treated clusters is small.

The problem is that the residuals must sum to zero over all the treated observations. Consider first the extreme case in which only the observations in cluster 1 are treated. In that case, $\sum_{i=1}^{N_1}\hat{u}_{1i} = 0$. This implies that the first term in expression (7) equals 0. The corresponding term on the right-hand side of equation (8) is certainly not 0. In fact, unless the proportion of treated observations is very large (which seems improbable when only one cluster is treated), the first term on the right-hand side of equation (8) will

[1] This particular factor is used by Stata and seems to have been introduced by them.

[2] In fact, with current computer hardware and software, it seems to become difficult to compute CV_2 once any of the N_g exceeds 5000 or so; see MacKinnon and Webb (2017*a*).

typically be larger than the second term, because $(1 - \bar{d})^2$ will typically be much larger than $\bar{d}^2$.

When two or more clusters are treated, the residuals for each treated cluster will not sum to zero, but they must sum to zero over all the treated clusters. Thus the sum of squared summations in the first term of (7) will always underestimate the corresponding triple summation in (8). As MacKinnon and Webb (2017*b*, Appendix A.3) shows, the underestimation should go away quite rapidly as G_1 increases. However, it is difficult to say just how large G_1 needs to be, even for the very simple model (5), because how well (7) estimates (8) depends on the sizes of the treated and untreated clusters and on the $\boldsymbol{\Omega}_g$ matrices. For more general models, it will also depend on the $\boldsymbol{X}_g$ matrices.

When the number of treated clusters is very small, it is quite possible for the CRVE (6) to underestimate the true variance of $\hat{\beta}_2$ by a factor of 25 or more, which implies that cluster-robust t statistics may be too large by a factor of five or more. This inevitably leads to confidence intervals that are much too narrow and tests that overreject very severely.

Bootstrap Methods

One widely-used way to obtain more reliable inferences than simply comparing a cluster-robust t statistic with the $t(G-1)$ distribution is to use a bootstrap test. Several different bootstrap tests are available, and they can produce very different results. As we will see in Section 5, bootstrap tests often yield considerably more reliable inferences than cluster-robust t tests, but they do not always work well. For a general introduction to bootstrap hypothesis testing, see Davidson and MacKinnon (2006).

When we perform a bootstrap test, we have to generate a large number of bootstrap samples. The alternative bootstrap methods that we consider differ only in how these bootstrap samples are generated. One particularly important issue concerns whether or not the bootstrap data-generating process, or DGP, imposes the null hypothesis. For concreteness, suppose we wish to test the hypothesis that β_k, the last element of $\boldsymbol{\beta}$, is zero. Then the bootstrap DGP could use either $\hat{\boldsymbol{\beta}}$, the unrestricted vector of OLS estimates, or $\tilde{\boldsymbol{\beta}}$, the vector with 0 as the last element and all other elements equal to the restricted OLS estimates.

For example, in the case of the pure treatment model (5), if the null hypothesis is that $\beta_2 = 0$, then $\tilde{\beta}_1 = \bar{y}$, the sample mean, and $\tilde{\beta}_2 = 0$. The estimate of β_1 under the null is just the sample mean in this simple case because, when $\beta_2 = 0$, equation (5) only contains a constant term. More generally, we would need to re-estimate the model subject to the restriction that $\beta_k = 0$.

It is good to choose B, the number of bootstrap samples, so that $\alpha(B+1)$ is an integer when α is the level of the test. Thus good values of B include 999 and 9999. However, if bootstrapping is very expensive, it may be possible to get away with a much smaller value of B by using a sequential procedure; see Davidson and MacKinnon (2000).

Assuming that B is fixed, the algorithm for computing a bootstrap P value for the hypothesis that $\beta_k = 0$ works as follows:

1. Estimate model (1) by OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$. Use these quantities to compute the cluster-robust t statistic $t_k = \hat{\beta}_k/\text{s.e.}(\hat{\beta}_k)$, where $\text{s.e.}(\hat{\beta}_k)$ denotes the square root of the k^{th} diagonal element of $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$.

2. Generate bootstrap samples $\boldsymbol{y}^{*b}$ for $b = 1, \ldots, B$, and re-estimate the model (1) to obtain $\hat{\boldsymbol{\beta}}^{*b}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{*b})$. The bootstrap samples may be generated in several different ways; see below.

3. For each bootstrap sample, compute the bootstrap t statistic, t_k^{*b}, as either

$$t_{kr}^{*b} = \frac{\hat{\beta}_k^{*b}}{\text{s.e.}(\hat{\beta}_k^{*b})} \quad \text{or} \quad t_{ku}^{*b} = \frac{\hat{\beta}_k^{*b} - \hat{\beta}_k}{\text{s.e.}(\hat{\beta}_k^{*b})}.$$

   If the bootstrap data generating process (DGP) imposes the null hypothesis, use t_{kr}^{*b} (the restricted version). If the bootstrap DGP does not impose the null hypothesis, use t_{ku}^{*b} (the unrestricted version).

4. Compute either the symmetric bootstrap P value

$$\hat{P}_{\text{S}}^{*} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\big(|t_k^{*b}| > |t_k|\big) \tag{9}$$

   or the equal-tail bootstrap P value

$$\hat{P}_{\text{ET}}^{*} = \frac{1}{B}\min\left(\sum_{b=1}^{B}\mathbb{I}(t_k^{*b} < t_k), \sum_{b=1}^{B}\mathbb{I}(t_k^{*b} \geq t_k)\right), \tag{10}$$

   where $\mathbb{I}(\cdot)$ denotes the indicator function, which equals 1 when its argument is true and 0 otherwise.

For many pure treatment and DiD models, we would expect t_k to be approximately symmetric around zero under the null hypothesis, so that equations (9) and (10) should yield very similar P values. Of course, this would generally not be true for dynamic models or models estimated by instrumental variables, since in both cases $\hat{\beta}_k$ might well be biased. In such cases, it would make sense to use $\hat{P}_{\text{ET}}^{*}$ rather than $\hat{P}_{\text{S}}^{*}$.

A number of different bootstrap methods can be used to generate the bootstrap samples, and, as we show in the next section, the finite-sample properties of these methods can differ enormously. The best-known method is probably the wild cluster bootstrap proposed in Cameron, Gelbach, and Miller (2008). For the restricted version of this procedure, we first re-estimate the model subject to the restriction(s)

to be tested, so as to obtain restricted estimates $\tilde{\boldsymbol{\beta}}$. The bootstrap DGP is then

$$\boldsymbol{y}_g^{*b} = \boldsymbol{X}_g\tilde{\boldsymbol{\beta}} + v_g^{*b}\tilde{\boldsymbol{u}}_g, \quad g = 1, \ldots, G, \tag{11}$$

where v_g^{*b} is a random variate with mean 0 and variance 1. A good choice in most cases is the Rademacher distribution, which takes the values 1 and -1 with equal probability; see Davidson and Flachaire (2008). Notice that, for each bootstrap sample, there is only one realization of this random variable for each cluster. In consequence, the bootstrap disturbances, $v_g^{*b}\tilde{\boldsymbol{u}}_g$, are independent across clusters but are supposed to mimic the covariance matrices of the residuals within each cluster.[3]

Combining the bootstrap DGP (11) with the algorithm given above yields a restricted wild cluster, or WCR, P value. If we replaced $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}_g$ in (11) by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}_g$, respectively, we would obtain an unrestricted wild cluster, or WCU, P value. Note that, in the latter case, we must use t_{ku}^{*b} rather than t_{kr}^{*b} for the bootstrap test statistics. Otherwise, the test would have no useful power. That is, the power of the test would be very similar to its size.

Recently, MacKinnon and Webb (2017a) suggested that it may be desirable to use the ordinary wild bootstrap instead of the wild cluster bootstrap for the model (5) when the number of treated clusters is small. The only difference between the restricted wild bootstrap (WR) and the restricted wild cluster bootstrap (WCR) is that the random variate v_g^{b*} in equation (11) is replaced by N_g random variates v_{gi}^{b*}, $i = 1, \ldots, N_g$. Since this eliminates the intracluster correlation that the wild cluster bootstrap is trying to capture, it may seem inappropriate. However, MacKinnon and Webb (2017a) shows that it can work very well in some cases, and Djogbenou, MacKinnon, and Nielsen (2017) proves that it is asymptotically valid. Of course, there is also an unrestricted wild bootstrap procedure (WU) which differs from WR in exactly the same way as WCU differs from WCR.

One important limitation of the wild and wild cluster bootstraps is that they can only be used with regression models. The models do not have to be linear, like (1), but they must have the form

$$y_{gi} = f(\boldsymbol{X}_{gi}, \boldsymbol{\beta}) + u_{gi}, \tag{12}$$

where $f(\boldsymbol{X}_{gi}, \boldsymbol{\beta})$ is a possibly nonlinear regression function that depends on a parameter vector $\boldsymbol{\beta}$ and a row vector of regressors $\boldsymbol{X}_{gi}$.

A very different way to generate bootstrap samples is to use the pairs cluster bootstrap, which was proposed (under a different name) in Bertrand, Duflo, and Mullainathan (2004). The idea of the pairs cluster bootstrap is to resample the entire $[\boldsymbol{y} \;\; \boldsymbol{X}]$ matrix by cluster. Thus each bootstrap sample consists of G submatrices chosen at random, with replacement, from the G submatrices $[\boldsymbol{y}_g \;\; \boldsymbol{X}_g]$ and stacked to form a matrix $[\boldsymbol{y}^{*b} \;\; \boldsymbol{X}^{*b}]$.

The pairs cluster bootstrap has one major advantage. It can be used for any sort of model in which the disturbances may be clustered, not just regression models. In particular, it can be used for binary response models such as the logit and probit models. However, it also has two serious disadvantages. The first is that, unless $N_g = N/G$ for all clusters, the bootstrap samples will almost never be the same size as the original sample. Some bootstrap samples will happen to contain a relatively large proportion of big clusters, and others will happen to contain a relatively large proportion of small clusters. When the N_g vary a lot, so will the sizes of the bootstrap samples. This will make it difficult for the distribution of the t_k^{*b} to mimic the distribution of t_k.

The second disadvantage of the pairs cluster bootstrap applies specifically to models with few treated clusters. We know from the results in MacKinnon and Webb (2017b) (and will also see in the next section) that the number of treated clusters, G_1, has an enormous impact on the rejection frequency of a cluster-robust t test. But G_1 necessarily varies when we use the pairs cluster bootstrap. Some bootstrap samples will contain more treated clusters than the actual sample, and some will contain fewer. Indeed, when G_1 is small, some bootstrap samples may well contain no treated clusters at all. This will happen for about 36.8% of them when $G_1 = 1$.[4] When it does happen, $\hat{\boldsymbol{\beta}}^{*b}$ cannot be computed, and the bootstrap sample has to be thrown out.

Another class of simulation-based procedures that has been proposed for making inferences about treatment effects at the cluster level is randomization inference, or RI. It was first suggested in this context by Conley and Taber (2011). Alternative RI procedures have been investigated by Canay, Romano, and Shaikh (2017), Ferman and Pinto (2015), and MacKinnon and Webb (2016). To keep this paper focused and reasonable in length, we do not consider any of these procedures in our simulations.

Simulation Experiments

In this section, we study the performance of cluster-robust t tests and five different bootstrap tests using simulation experiments. We study both the pure treatment model (5) and a DiD regression model. Our focus is on the number of treated clusters, G_1, holding N and G fixed. We report

[3] When the number of clusters is small, the fact that a Rademacher random variate can take on only two values means that the number of possible bootstrap samples, which is 2^G, is rather small. For $G \leq 12$, it may be better to use another discrete distribution which can take on six values; see Webb (2014).

[4] The probability that any given cluster will not appear in a particular bootstrap sample is $\big((G-1)/G\big)^G$. This converges to $\exp(-1) = 1/(2.71828) = 0.36788$ as $G \to \infty$.

all our results graphically, with 5% rejection frequencies[5] on the vertical axis and G_1 on the horizontal axis.

The first set of experiments is for the pure treatment model (5). The value of G is either 12 or 24, and we set $N = 100G$. We would have obtained extremely similar results for most methods if N had been 5, 10, or even 100 times larger. The disturbances, u_{gi}, are normally distributed and generated by a random effects model with intra-cluster correlation coefficient $\rho_g = 0.05$. We do not study the role of ρ_g in detail because previous research (along with some experiments that we do not report) has shown that it typically has a very modest effect for the WCR and WCU bootstraps. For obvious reasons, it does have a somewhat larger effect for the WR and WU bootstraps, however.

Results in MacKinnon and Webb (2017*b*) show that the amount of variation in cluster sizes can be very important. In order to allow for possibly unbalanced cluster sizes, N_g is determined by a parameter γ, as follows:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \qquad (13)$$

where $[\cdot]$ denotes the integer part of its argument, and $N_G = N - \sum_{j=1}^{G-1} N_g$. Every N_g is equal to $N/G = 100$ when $\gamma = 0$. As γ increases or decreases, cluster sizes become increasingly unbalanced. For $\gamma > 0$, the N_g increase as g increases, and for $\gamma < 0$, they decrease.

All experiments have 400,000 replications. We use such a large number because rejection frequencies often differ very little among alternative methods, and it would otherwise be difficult to distinguish between systematic differences and experimental errors. All bootstrap methods use $B = 399$.

Panels a) and b) of Figure 1 show results for experiments with $G = 12$. In panel a), $\gamma = 0$, so that $N_g = 100$ for all g. In panel b), $\gamma = 2$, so that N_g increases from 34 to 217 as g increases from 1 to 12. The clusters are always treated in increasing order. Thus, if $G_1 = 2$ and $\gamma > 0$, only the smallest and second-smallest clusters are treated. This is probably not very realistic, but it illustrates the importance of differing cluster sizes. In practice, even if cluster sizes varied a lot, it would probably not often be the case that only the very smallest (or very largest) clusters were treated. Thus the results in panel b) are probably for an extreme case.

All the results in panel a) of Figure 1 are symmetric around $G_1 = 6$. This reflects the fact that, for constant cluster sizes, there is no real difference between G_1 and G_0. It is evident that the cluster-robust t statistic always overrejects, and it does so very severely for $G_1 \leq 2$ and $G_1 \geq 10$. The pairs cluster and WCU bootstraps also overreject very severely for $G_1 = 1$ and $G_1 = 11$, but they improve more rapidly as G_1 becomes less extreme, and the latter actually performs very well for $4 \leq G_1 \leq 8$. In contrast, the pairs bootstrap underrejects quite severely for those cases.

The WCR bootstrap underrejects extremely severely for $G_1 = 1$ and $G_1 = 11$ (the actual rejection rates are about 1 in 10,000) and very severely for $G_1 = 2$ and $G_1 = 10$. However, it works quite well for $3 \leq G_1 \leq 9$. The performance of the cluster-robust t statistic and the WCR and WCU bootstraps here are precisely what the theoretical results in MacKinnon and Webb (2017*b*, Section 6) predict.

To our knowledge, the performance of the pairs cluster bootstrap for treatment models has not been studied. However, it could have been predicted from the results of Davidson and MacKinnon (1999). That paper studies the size distortion of parametric bootstrap tests, and the value of G_1 here plays essentially the same role as the parameters in that paper. The paper shows that whether a bootstrap test overrejects or underrejects is determined by how much the distributions of the bootstrap test statistics depend on the parameter estimates and on how biased and noisy those estimates are.

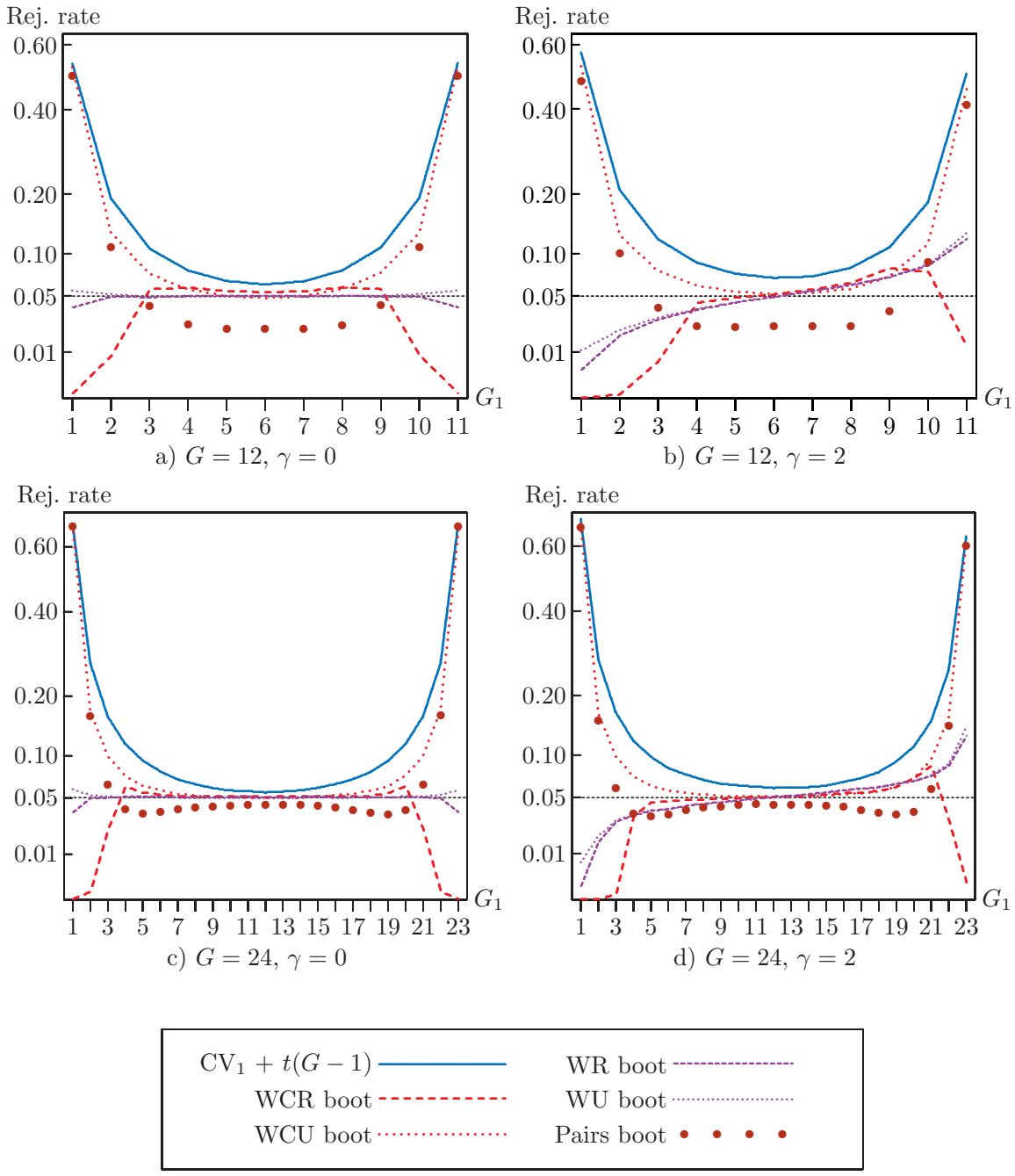
In the case of Figure 1, many of the errors in rejection frequency for the pairs cluster bootstrap arise from the fact that the bootstrap samples contain different groups. Imagine a sample with four groups denoted A, B, C, and D, of which only A is treated, so that $G_1 = 1$. If we call the number of treated clusters in a bootstrap sample G_1^*, then the bootstrap sample $\{A, B, A, C\}$ would have $G_1^* = 2$, whereas the sample $\{B, D, B, C\}$ would have $G_1^* = 0$. When $G_1^* = 0$, we obviously cannot estimate the model, so any samples that do not contain A would have to be discarded. Thus all the bootstrap samples that we can actually use have $G_1^* \geq 1$ (and also $G_0^* \leq 3$). On average, the values of G_1^* must be greater than G_1 whenever $G_1 = 1$.

The same thing happens in our experiments. When G_1 is very small or very large, many of the bootstrap samples have values of G_1^* that are less extreme than G_1 itself. This means that the actual t statistic t_k is being compared with a mixture of distributions, many of which involve less extreme values of G_1^*. In contrast, when G_1 takes on values between 4 and 8, the bootstrap t statistics with which t_k is being compared are often associated with substantially more extreme values of G_1^*. In the former case, we get severe overrejection, and in the latter case quite severe underrejection.

The WR and WU bootstraps work extraordinarily well in panel a) of Figure 1. They perform well even for extreme values of G_1, and they differ noticeably from each other only for $G_1 = 1$ and $G_1 = 11$. Note that they would have differed more for $\rho_g > 0.05$ and even less for $\rho_g < 0.05$. This is precisely what the results of MacKinnon and Webb (2017*a*)

[5] A 5% rejection frequency is the proportion of the P values over the 400,000 replications that are below 0.05. If the null is true and the procedure works perfectly, then the 5% rejection frequency should be 0.05, plus or minus a little bit of experimental error. When the true rejection frequency is 0.05, the standard error of the estimate is 0.00034. Thus we would expect to see numbers between 0.04932 and 0.05068 about 95% of the time for procedures that work perfectly.

Figure 1: Rejection Frequencies for Pure Treatment Model



In all panels, the number of treated clusters (G_1) is on the horizontal axis.

predict. Unfortunately, those results require that all clusters be the same size and have the same patterns of intra-cluster correlation, which is the case in panel a) of the figure.

In panel b) of Figure 1, the variation in cluster sizes causes a number of asymmetries. This is evident for all the methods, but particularly for WCR, WR, and WU. In the case of WCR, there is now underrejection for $G_1 = 3$ as well as $G_1 \leq 2$, and there is quite substantial overrejection for $G_1 = 9$ and $G_1 = 10$. The two wild bootstrap methods, WR and WU, both underreject for small values of G_1 and overreject for large values, as predicted in MacKinnon and Webb (2017*a*). However, they still perform better than WCR and WCU for the two largest and three smallest values of G_1.

Panels c) and d) of Figure 1 are similar to panels a) and b), respectively, except that $G = 24$ and G_1 runs from 1 to 23. Many of the results do not change much when we double the number of clusters. However, it is noteworthy that rejection rates for the t statistic, the WCU bootstrap, and the pairs cluster bootstrap are actually worse for $G_1 = 1$ and $G_0 = 1$ than they were before. On the other hand, all methods now work better for intermediate values of G_1, with WCR and WCU performing very well for $8 \leq G_1 \leq 16$.

The pairs cluster bootstrap underrejects for all but the three most extreme values of G_1 and G_0, but it does so much less severely than in the top two panels. The latter could have been predicted from the analysis above. The range of values of G_1 for which the rejection rates of the t statistic do not change very much is now much wider. This means that these rates do not vary as much across the G_1^* for the various bootstrap samples as they did when $G = 12$, so that the differences between the actual G_1 for t_k and the various G_1^* for the corresponding t_k^{b*} do not matter as much.

Overall, when we compare the results in the bottom two panels with those in the top two, it appears that increasing the number of clusters improves the performance of all methods, but only if G_1/G is not too large or small. This is precisely what would be expected from the asymptotic theory in Djogbenou, MacKinnon, and Nielsen (2017), where regularity conditions effectively require the number of treated clusters to rise with (but not necessarily as fast as) the sample size.

The second set of experiments is for a DiD regression model. The model that we estimate can be written as

$$y_{gi} = \sum_{t=1}^{T} \delta_t D_{gi}^t + \beta_2 d_{gi} + u_{gi}, \qquad (14)$$
$$i = 1 \ldots, N_g, \quad g = 1, \ldots, G.$$

The constant term in (5) has been replaced by T dummy variables D_{gi}^t, each of which takes the value 1 for observations associated with year t and 0 otherwise. The treatment variable d_{gi} now takes the value 1 only for treated clusters during the years in which they are treated.

In practice, investigators would often add $G - 1$ dummy variables for all but one of the clusters to regression (14). We did not do so because it would have made the experiments more expensive, and the cluster dummies would have simply offset the cluster-specific shocks in the random effects specification of the u_{gi}.[6] We choose which clusters are to be treated and then allow our program to decide at random when treatment is to commence. We set $T = 20$ and allow treatment to start as early as year 6 and as late as year 16.

The big difference between the treatment model (5) and the DiD model (14) is that, for the latter, there is no symmetry between G_1 and G_0. In fact, it is entirely possible to have $G_1 = G$, so that $G_0 = 0$. In that case, identification comes from the fact that all of the treated clusters contain some untreated observations. We therefore expect results for small values of G_1 to resemble the ones in Figure 1 but results for large values to be very different, and that is indeed what we see in Figure 2.

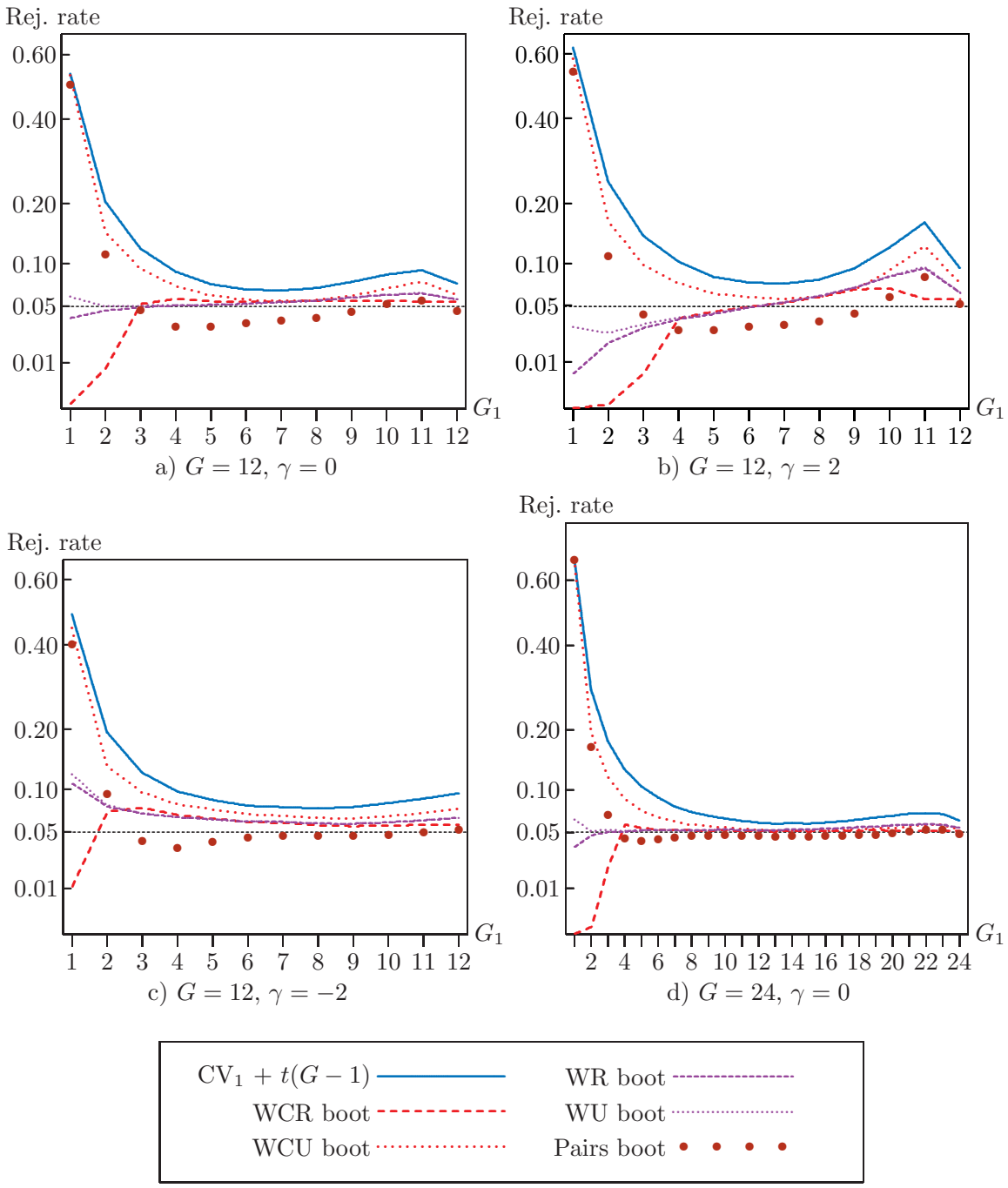
The top two panels of Figure 2 are comparable to the same two panels of Figure 1. Notice that G_1 runs from 1 to 12 instead of from 1 to 11. In both cases, results for small values of G_1 are quite similar across the two figures. There is severe overrejection for the t statistic, the WCU bootstrap, and the pairs bootstrap, together with severe underrejection for the WCR bootstrap. However, the results for other values of G_1 are not very similar across the two figures. In panel a), all the wild bootstrap methods work fairly well for intermediate and large values of G_1, and even the pairs cluster bootstrap performs much better than it did before. Perhaps surprisingly, the results for $G_1 = 12$ are noticeably better than for $G_1 = 11$. This probably happens because $G_0 = 1$ when $G_1 = 11$, and having just one non-treated cluster tends to cause problems. In contrast, when $G_1 = 12$, there are no longer any non-treated clusters, just 12 clusters with various proportions of treated observations.

Panel c) of Figure 2 deals with the case of $G = 12$ and $\gamma = -2$. The variation in cluster sizes here is very similar to the variation in panel b), but now the largest clusters are treated first. There are some very substantial differences between panels b) and c). In the latter, WR and WU always overreject, especially for small values of G_1, and WCR overrejects for $G_1 \geq 2$. Perhaps surprisingly, the pairs bootstrap is actually the best method for $G_1 \geq 5$.

Finally, panel d) of Figure 2 deals with the case of $G = 24$ and $\gamma = 0$. All the bootstrap methods work quite well for $G_1 \geq 4$, and WCR works extremely well for $G_1 \geq 7$. It is difficult to see in the figure, but WCU and the two ordinary

[6] This does not mean that adding group-specific fixed effects solves the problem of intracluster correlation; it would merely do so for our specific DGP. In the "placebo law" experiments of Bertrand, Duflo, and Mullainathan (2004) and MacKinnon and Webb (2017*b*), which use real data from the Current Population Survey, standard errors that are robust to heteroskedasticity but not to intracluster correlation yield extremely severe errors of inference despite the presence of group-specific fixed effects.

Figure 2: Rejection Frequencies for Difference-in-Differences Model



In all panels, the number of treated clusters (G_1) is on the horizontal axis.

wild bootstrap methods overreject slightly for the largest values of G_1. The latter work quite well even for small values of G_1, but this undoubtedly reflects the rather simple experimental design.

We remark that all the results for DiD models would have been different if the "years" in which treatment begins had been different. They might have been quite different if we had conditioned on a particular set of years for treatment rather than choosing them at random. With our procedure, results tend to average out across replications.

Empirical Example

In this section, we consider an empirical example from Burden, Canon, Mayer, and Moynihan (2017). It uses county-level data to analyze the effects of certain state-level voting laws, particularly early voting, on Democratic vote share. In a portion of the analysis, the authors attempt to estimate the following non-panel DiD model to analyze the effect of early voting laws:

$$\begin{aligned}\text{demdiff}_{cs} = \beta_0 + \beta_1 \text{EV}_{cs} + \beta_2 \texttt{felon}_{cs} \\ + \beta_3 \texttt{id}_{cs} + \boldsymbol{X}_{cs}\boldsymbol{\beta}_4 + \epsilon_{cs}.\end{aligned} \tag{15}$$

Here demdiff_{cs} represents the difference in Democratic vote share between the 2008 and 2012 elections, for county c in state s. β_1 is the coefficient of interest. EV takes on the values -1, 0, or 1, if a state either repealed, did not change, or adopted early voting laws, respectively. Between 2008 and 2012, California and Maryland adopted, while New Jersey repealed, their early voting laws. This specification assumes a symmetric treatment effect for repealing or adopting; we will later re-estimate this model using asymmetric treatment effects. Additionally, `felon` and `id` are binary variables representing changes in felon disenfranchisement and ID requirement laws, and $\boldsymbol{X}$ represents various county-level demographic covariates. All estimates are weighted by county population. The U.S. has a total of 3,144 counties and county equivalents. The dataset has $N = 3,112$ usable observations collected from every state including D.C. but excluding Alaska, so that $G = 50$. Cluster sizes range from 1 county in D.C. to 254 counties in Texas.

We attempt to reproduce Column 4 from Table 7 of Burden et al. (2017), which contains their difference-in-differences analysis comparing the 2008 and 2012 elections. While attempting to reproduce the results, two problems were found in the replication files provided. Firstly, three control counties were mistakenly coded as treated: one in Alaska, one in Colorado, and one in DC. Additionally, California, which adopted early voting, was mistakenly coded as a control. We estimate the model using the data as coded in the 'miscoded' column, and then re-estimate it using the proper coding in the 'corrected' column.

Secondly, standard errors were clustered by county Federal Information Processing Standards (fips). County fips codes are unique within but not across states. The prevailing commonality amongst these counties is their position alphabetically within state. Clustering by fips assumes that there is potential correlation across the alphabetically first, second, third, etc. clusters across states, but no correlation within states. It appears that the authors intended to cluster at the individual county level; however, since observations are at the county level, each cluster would only contain one observation in that case. Clustering by county would then result in heteroskedasticity-robust rather than cluster-robust standard errors. Since earlier results in Burden et al. (2017) were clustered by state, and laws change at the state level, it seems sensible to cluster by state here as well.

Table 1 demonstrates the limitations of inference when few groups are treated. The first two columns seek to estimate the model used in Burden et al. (2017). The other two columns show this model without the symmetry assumption, where adopting states (CA, MD) and repealing states (NJ) are considered separately. The top row of Table 1 reports the estimate of β_1 or its asymmetric equivalent.[7]

The second panel reports three asymptotic P values based on robust, CRVE-fips, and CRVE-state standard errors. The bottom panel reports two bootstrap P values (WCR and WCU), both of which are calculated using $B = 99,999$ and are clustered by state.

In column 1, we replicate the analysis of Burden et al. (2017). The Robust and CRVE-fips P values are highly significant. However, clustering by state greatly reduces the evidence against the null, and the bootstrap P values are highly insignificant. In column 2, where treatment status has been corrected, all of the P values are insignificant.

When we estimate the asymmetric effects, we can really see the problems of few treated clusters, especially in evaluating the effects of a repeal of early voting where $G_1 = 1$. Here we observe that all of the asymptotic and WCU P values are highly significant, whereas the WCR P value is highly insignificant. This is exactly what the theory predicts and what the simulation results in Figure 2 show.

[7] For the asymmetric equivalent, rather than β_1 being the coefficient when EV_{cs} is equal to -1 or $+1$, we in effect estimate two coefficients. The first, β_1^a, is for states that adopted early voting, so that $\text{EV}_{cs} = -1$, and the second, β_1^r, is for states that repealed early voting, so that $\text{EV}_{cs} = 1$. For the asymmetric estimates, we have two binary variables, one for repeal, which is equal to 1 for New Jersey and equal to 0 for all other states, and one for adopt, which is equal to 1 for California and Maryland and equal to 0 for all other states. This explains why the coefficient is negative in the "corrected" column, even though both coefficients are positive in the asymmetric estimates. New Jersey saw its Democratic vote share increase from 2008 to 2012, despite eliminating early voting. The coding of $\text{EV}_{cs} = -1$ essentially forces the coefficient to be negative to pick up the large increase that occurred in New Jersey, despite the fact that Maryland and California also had (difference-in-differences) increases after adopting early voting.

Table 1: Effects of Early Voting Laws on Democratic Vote Shares

	Symmetric Effect		Asymmetric Effect	
	Miscoded	Corrected	EV Adopted	EV Repealed
$\hat{\beta}_1$	−1.458	−0.254	0.700	2.795
Robust	0.000	0.401	0.041	0.000
CRVE (Ctyfips)	0.000	0.325	0.055	0.000
CRVE (State)	0.088	0.722	0.144	0.000
WCU	0.500	0.799	0.208	0.000
WCR	0.740	0.880	0.300	0.423
Treated States	-	3	2	1

Empirical example from Burden, Canon, Mayer, and Moynihan (2017). Treated states in 'Miscoded' column omitted due to coding issues in treatment assignment. All entries except the estimates of β_1 are P values. Estimates are weighted by county population. WCU and WCR P values are clustered at the state level and based on 99,999 bootstraps.

While the example above highlights the difficulties of statistical inference when there are few treated groups, we do not regard these findings as challenging the conclusions in Burden et al. (2017). In fact, when we evaluate the asymmetric estimates, we see that the states which adopted early voting had a statistically insignificant increase in the Democratic vote share, while the state that repealed early voting had a statistically ambiguous *increase* in that share. Taken together, these results present evidence against the conventional wisdom that early voting laws favor Democrats.

Conclusions

When a regression model is used to estimate treatment effects, cluster-robust standard errors can be extremely misleading when cluster sizes vary a lot and/or when the number of treated clusters (G_1) is small. This is true for both pure treatment models and difference-in-differences (DiD) models. One simple way to see whether there is a problem is to calculate bootstrap P values using two variants of the wild cluster bootstrap. If the WCR (restricted) and WCU (unrestricted) bootstraps yield very different inferences, then there is definitely a problem. Normally, in such cases, WCU will reject and WCR will fail to reject. Other methods may or may not yield reliable results.

Unfortunately, even when WCR and WCU roughly agree, there may be a problem; consider the results for $G_1 = 10$ in panel b) of Figure 1. However, based on the simulation results here and in MacKinnon and Webb (2017*a*,*b*), agreement between WCR and WCU seems to rule out really severe errors of inference. These are often associated with very small values of G_1, where WCR and WCU tend to disagree sharply.

It is also worth trying other bootstrap methods. When cluster sizes are similar, the ordinary wild bootstrap can work surprisingly well, even when G_1 is very small, but the pairs cluster bootstrap typically overrejects in that case. The latter is rarely the procedure of choice for regression models, but it can be useful for nonlinear models such as logit and probit. However, it tends to overreject severely when G_1 is small, and it can underreject severely when the number of clusters, G, is small and G_1/G is not close to 0 or 1.

There is an important difference between pure treatment models and DiD models. For the former, small numbers of untreated clusters are just as bad as small numbers of treated ones. For the latter, reasonably reliable results can be obtained even when all clusters are treated, provided treatment starts at different times for different clusters.

Acknowledgments

We are grateful to Justin Esarey for several very helpful suggestions and to Joshua Roxborough for valuable research assistance. This research was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada. Some of the computations were performed at the Centre for Advanced Computing at Queen's University.

References

Angrist, Joshua D. and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* First ed. Princeton: Princeton University Press.

Bell, Robert M. and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with

Multi-Stage Samples." *Survey Methodology* 28:169–181.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119:249–275.

Bester, C. Alan, Timothy G. Conley and Christian B. Hansen. 2011. "Inference with Dependent Data Using Cluster Covariance Estimators." *Journal of Econometrics* 165:137–151.

Burden, Barry C., David T. Canon, Kenneth R. Mayer and Donald P. Moynihan. 2017. "The Complicated Partisan Effects of State Election Laws." *Political Research Quarterly* 70:564–576.

Cameron, A. Colin and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster Robust Inference." *Journal of Human Resources* 50:317–372.

Cameron, A. Colin, Jonah B. Gelbach and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90:414–427.

Canay, Ivan A., Joseph P. Romano and Azeem M. Shaikh. 2017. "Randomization Tests under an Approximate Symmetry Assumption." *Econometrica* 85:1013–1030.

Carter, Andrew V., Kevin T. Schnepel and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity." *Review of Economics and Statistics* 99, Forthcoming.

Conley, Timothy G. and Christopher R. Taber. 2011. "Inference with 'Difference in Differences' with a Small Number of Policy Changes." *Review of Economics and Statistics* 93:113–125.

Davidson, Russell and Emmanuel Flachaire. 2008. "The Wild Bootstrap, Tamed at Last." *Journal of Econometrics* 146:162–169.

Davidson, Russell and James G. MacKinnon. 1999. "The Size Distortion of Bootstrap Tests." *Econometric Theory* 15:361–376.

Davidson, Russell and James G. MacKinnon. 2000. "Bootstrap Tests: How Many Bootstraps?" *Econometric Reviews* 19:55–68.

Davidson, Russell and James G. MacKinnon. 2006. "Bootstrap Methods in Econometrics." In *Palgrave Handbooks of Econometrics: Volume 1 Econometric Theory*, ed. Terence C. Mills and Kerry D. Patterson. Palgrave pp. 812–838.

Djogbenou, Antoine, James G. MacKinnon and Morten Ø Nielsen. 2017. "Bootstrap Inference with Clustered Errors." QED working paper Queen's University.

Esarey, Justin and Andrew Menger. 2017. "Practical and Effective Approaches to Dealing with Clustered Data." *Political Science Research and Methods* 5, Forthcoming.

Ferman, Bruno and Christine Pinto. 2015. "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity." Technical report Sao Paulo School of Economics.

Imbens, Guido W. and Michal Kolesár. 2016. "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics* 98:701–712.

MacKinnon, James G. 2012. "Thirty Years of Heteroskedasticity-Robust Inference." In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. Xaiohong Chen and Norman R. Swanson. Springer pp. 437–461.

MacKinnon, James G. and Halbert White. 1985. "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics* 29:305–325.

MacKinnon, James G. and Matthew D. Webb. 2016. "Randomization Inference for Difference- in-Differences with Few Treated Clusters." QED Working Paper 1355 Queen's University. URL: http://ideas.repec.org/p/qed/wpaper/1355.html.

MacKinnon, James G. and Matthew D. Webb. 2017*a*. "The Subcluster Wild Bootstrap for Few (Treated) Clusters." QED Working Paper 1364 Queen's University.

MacKinnon, James G. and Matthew D. Webb. 2017*b*. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32:233–254.

MacKinnon, James G., Morten Ø. Nielsen and Matthew D. Webb. 2017. "Bootstrap and Asymptotic Inference with Multiway Clustering." QED Working Paper 1386 Queen's University.

Pustejovsky, James E. and Elizabeth Tipton. 2017. "Small Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." *Journal of Business and Economic Statistics* 35, Forthcom-

ing.

Webb, Matthew D. 2014. "Reworking Wild Bootstrap Based Inference for Clustered Errors." QED Working Paper 1315 Queen's University. URL: http://ideas.repec.org/p/qed/wpaper/1298.html.

Young, Alwyn. 2016. "Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections." Technical report London School of Economics.

Response to MacKinnon and Webb

Barry C. Burden
University of Wisconsin-Madison
bcburden@wisc.edu

David T. Canon
University of Wisconsin-Madison
dcanon@polisci.wisc.edu

Kenneth R. Mayer
University of Wisconsin-Madison
krmayer@wisc.edu

Donald P. Moynihan
University of Wisconsin-Madison
dmoynihan@lafollette.wisc.edu

MacKinnon and Webb offer a useful analysis of how the uncertainty of causal effects can be underestimated when observations are clustered and the treatment is applied to a very large or vary small share of the clusters. Their mathematical exposition, simulation exercises, and replication analysis provide a helpful guide for how to proceed when data are poorly behaved in this way. These are valuable lessons for researchers studying impacts of policy in observational data where policies tend to be sluggish and thus do not generate much variability in the key explanatory variables.

Correction of Two Errors

MacKinnon and Webb find two errors in our analysis, while nonetheless concluding "we do not regard these findings as challenging the conclusions of Burden et al. (2017)." Although we are embarrassed by the mistakes, we are also grateful for their discovery. Our commitment to transparency is reflected by the fact the data was been made public for replication purposes since well before the article was published. We have posted corrected versions of the replication files and published a corrigendum with the journal where the article was original published.

Fortunately, none of the other analyses in our article were affected. It is only Table 7 where errors affect the analysis. Tables 2 through 6 remain intact.

We concede that when corrections are made the effect of early voting drops from statistical significance in the model of the difference in the Democratic vote between 2008 and 2012. All of the various standard errors they report are far too large to reject the null hypothesis.

The Problem of Limited Variation

The episode highlights the tradeoffs that researchers face between applying what appears to be a theoretically superior estimation technique (i.e., difference-in-difference) and the practical constraints of a particular application (i.e., limited variation in treatment variables) that make its use intractable. In the case of our analysis, election laws do not change rapidly, and the conclusions of our analysis were largely based on cross-sectional analyses (Tables 2–6), with the difference-in-difference largely offered as a supplemental analysis.

We are in agreement with MacKinnon and Webb that models designed to estimate causal effects (or even simple relationships) may be quite tenuous when the number of clusters is small and the clusters are treated in a highly unbalanced fashion. In fact, we explained our reluctance to apply the difference-in-difference model to our data because of the limited leverage available. We were explicit about our reservations in this regard. As our article stated:

"A limitation of the difference-in-difference approach in our application is that few states actually changed their election laws between elections. As Table A1 (see Supplemental Material) shows, for some combinations of laws there are no changes at all. For others, the number of states changing is as low as one or two. As result, we cannot include some of the variables in the model because they do not change. For some other variables, the interpretation of the coefficients would be ambiguous given the small number of states involved; the dummy variables essentially become fixed effects for one or two states" (p. 572).

This is unfortunate in our application because the difference-in-difference models are likely to be viewed as more convincing than the cross-sectional models. This is why we offered theory suggesting that the more robust cross-sectional results were not likely to suffer from endogeneity.

The null result in the difference-in-difference models is not especially surprising given our warning above about the limited leverage provided by the dataset. Indeed, the same variable was insignificant in our model of the Democratic vote between 2004 and 2008 that we also reported in Table 7. We are left to conclude that the data are not amenable to detecting effects using difference-in-difference models. Perhaps researchers will collect data from more elections to provide more variation in the key variable and estimate param-

eters more efficiently.

In addition to simply replicating our analysis, MacKinnon and Webb also conduct an extension to explore asymmetric effects. They separate the treated states into those where early voting was adopted and where early voting was repealed. We agree that researchers ought to investigate such asymmetries. We recommended as much in our article: "As early voting is being rolled back in some states, future research should explore the potential asymmetry between the expansion and contraction of election practices" (p. 573). However, we think this is not feasible with existing data. As MacKinnon and Webb note, only two states adopted early voting and only one state repealed early voting. As a result, analyzing these cases separately as they do essentially renders the treatment variables to be little more than fixed effects for one or two states, as we warned in our article. The coefficients might be statistically significant using various standard error calculations, but it is not clear that MacKinnon and Webb are actually estimating the treatment effects rather than something idiosyncratic about one or two states.

Conclusion

While the errors made in our difference-in-difference analysis were regrettable, we think the greater lesson from the skilled analysis of MacKinnon and Webb is to raise further doubt about whether this tool is simply unsuitable in such a policy setting. While all else is equal, it may offer a superior mode of analysis; but all else is not equal. Researchers need to find the best mode of analysis to fit with the limitations of the data.

Rice University
Department of Political Science
Herzstein Hall
6100 Main Street (P.O. Box 1892)
Houston, Texas 77251-1892

The Political Methodologist is the newsletter of the Political Methodology Section of the American Political Science Association. The support of the Department of Political Science at Rice University in helping to defray the editorial and production costs of the newsletter is gratefully acknowledged.

Subscriptions to *TPM* are free for members of the APSA's Methodology Section. Please contact APSA (202-483-2512) if you are interested in joining the section. Dues are $29.00 per year for students and $44.00 per year for other members. They include a free subscription to *Political Analysis*, the quarterly journal of the section.

Submissions to *TPM* are always welcome. Articles may be sent to any of the editors, by e-mail if possible. Alternatively, submissions can be made on diskette as plain ASCII files sent to Justin Esarey, 108 Herzstein Hall, 6100 Main Street (P.O. Box 1892), Houston, TX 77251-1892. LaTeX format files are especially encouraged.

TPM was produced using LaTeX.

THE SOCIETY FOR POLITICAL METHODOLOGY

ad indicia spectate

Corrigendum to "Lowering the Threshold of Statistical Significance to p $<$ 0.005 to Encourage Enriched Theories of Politics" and "Questions and Answers: Reproducibility and a Stricter Threshold for Statistical Significance"

Justin Esarey
Wake Forest University
Department of Politics and International Affairs
justin@justinesarey.com

December 1, 2018

Although *The Political Methodologist* is a newsletter and blog, not a peer-reviewed publication, I still think it's important for us to recognize and correct substantively important errors. In this case, I'm sad to report such errors in two things I wrote for *TPM*. The error is the same in both cases.

In "Lowering the Threshold of Statistical Significance to p $<$ 0.005 to Encourage Enriched Theories of Politics," I claimed that:

> When K-many statistically independent tests are performed on pre-specified hypotheses that must be jointly confirmed in order to support a theory, the chance of simultaneously rejecting them all by chance is α^K where $p < \alpha$ is the critical condition for statistical significance in an individual test. As K increases, the α value for each individual study can fall and the overall power of the study often (though not always) increases.

This argument is offered to support the conclusion that "moving the threshold for statistical significance from $\alpha = 0.05$ to $\alpha = 0.005$ would benefit political science if we adapt to this reform by developing richer, more robust theories that admit multiple predictions."

Similarly, in "Questions and Answers: Reproducibility and a Stricter Threshold for Statistical Significance," I claimed that:

> Another measure to lower Type I error (and the one that I discuss in my article in *The Political Methodologist*) is to pre-specify a larger number of different hypotheses from a theory and to jointly test these hypotheses. Because the probability of simultaneously confirming multiple disparate predictions by chance is (almost always) lower than the probability of singly confirming one of them, the size of each individual test can be larger than the overall size of the test, allowing for the possibility that the overall test is substantially more powerful at a given size.

This reasoning, which is similar to reasoning offered in Esarey and Sumner (2018*b*), is incorrect; it would only be true when all predicted parameters were equal to zero. When the alternative hypothesis is that multiple directional predictions for parameters, for example $\beta_i > 0$ for $i \in 1...K$, separate t-tests rejecting each individual null ($\beta_i \leq 0$) separately using t-tests with size α will jointly reject all the null hypotheses at most α proportion of the time. The key insight is that the joint null hypothesis space includes the possibility that some β_i parameters match the predictions while others do not; if (for example) $\beta_1 = 0$ and all other $\beta_{i \neq 1}$ are very large, the probability of falsely rejecting the joint null hypothesis is the α for the test of β_1. As we note in Esarey and Sumner (2018*a*), this is discussed and proved in Silvapulle and Sen (2005, Section 5.3), especially in proposition 5.3.1, and in Casella and Berger 2002, Section 8.2.3 and 8.3.3. Silvapulle and Sen cite Lehmann (1952); Berger (1982); Cohen, Gatsonis and Marden (1983); and Berger (1997) (among others) as sources for this argument. Associated calculations (such as that in Figure 4 of "Lowering the Threshold of

Statistical Significance to p < 0.005 to Encourage Enriched Theories of Politics") are based on the same error and therefore incorrect.

The upshot is that my argument for making additional theoretical predictions in order to facilitate lowering the threshold for statistical significance to $\alpha = 0.005$ is based on faulty reasoning and incorrect.

I plan to post this correction as an addendum to both of the print editions featuring these articles.

References

Berger, Roger L. 1982. "Multiparameter Hypothesis Testing and Acceptance Sampling." *Technometrics* 24(4):295–300.

Berger, Roger L. 1997. Likelihood ratio tests and intersection-union tests. In *Advances in statistical decision theory and applications*, ed. Subramanian Panchapakesan and Narayanaswamy Balakrishnan. Boston: Birkhäuser pp. 225–237.

Casella, George and Roger L.. Berger. 2002. *Statistical Inference, Second Edition*. Belmont, CA: Brooks/Cole.

Cohen, Arthur, Constantine Gatsonis and John I. Marden. 1983. "Hypothesis testing for marginal probabilities in a 2 × 2 × 2 contingency table with conditional independence." *Journal of the American Statistical Association* 78(384):920–929.

Esarey, Justin and Jane Lawrence Sumner. 2018*a*. "Corrigendum to Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." Online. URL: `http://justinesarey.com/interaction-overconfidence-corrigendum.pdf`.

Esarey, Justin and Jane Lawrence Sumner. 2018*b*. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." *Comparative Political Studies* 51(9):1144–1176. DOI: `https://doi.org/10.1177/0010414017730080`.

Lehmann, Erich L. 1952. "Testing multiparameter hypotheses." *The Annals of Mathematical Statistics* pp. 541–552.

Silvapulle, Mervyn J. and Pranab K. Sen. 2005. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Hoboken, NJ: Wiley.