How Do We Know What We Know? Learning from Monte Carlo Simulations

Vincent Hopkins, University of British Columbia
Ali Kagalwala, Texas A&M University
Andrew Q. Philips, University of Colorado Boulder
Mark Pickup, Simon Fraser University
Guy D. Whitten, Texas A&M University

Monte Carlo simulations are commonly used to test the performance of estimators and models from rival methods, under a range of data-generating processes. This tool improves our understanding of the relative merits of rival methods in different contexts, such as varying sample sizes and violations of assumptions. When used, it is common to report the bias or the root mean squared error of the different methods. It is far less common to report the standard deviation, overconfidence, coverage probability, or power. Each of these six performance statistics provides important, and often differing, information regarding a method's performance. Here, we present a structured way to think about Monte Carlo performance statistics. In replications of three prominent papers, we demonstrate the utility of our approach and provide new substantive results about the performance of rival methods.

One of the great strengths of political science as a discipline has been our enthusiasm for embracing new methods for testing hypotheses. Whenever the use of a new method is proposed, one of the first questions that researchers ask is how it performs relative to existing methods. To make such assessments, researchers have relied heavily on performance statistics (e.g., root mean squared error [RMSE]) of estimators or models from rival methods in Monte Carlo simulations. This approach of comparing rival methods has become pervasive in political methodology and is a core component of some of the most highly cited papers in all of political science (e.g., Beck and Katz 1995; Keele and Kelly 2006; King and Zeng 2001; Plümper and Troeger 2007).

While papers taking this approach have provided a wealth of helpful advice to applied researchers, we argue that this advice has often been based on too little information. As we demonstrate in our review of the literature below, many papers that use Monte Carlo simulations to make comparisons between rival methods use only one or two performance statistics and rely most heavily on measures of bias and RMSE. While these are excellent criteria for assessing relative performance, we argue that other easily calculable performance statistics such as standard deviation (SD), overconfidence, coverage, and power often should also be reported. Doing so will allow researchers to make more informed decisions about which methods are preferred under different circumstances.

We write for two audiences: those who wish to produce Monte Carlo simulations to examine the relative performance of different methods, and those who wish to read the results of Monte Carlo simulations to learn about the relative performance of different methods. For the first group, we

Vincent Hopkins (vince.hopkins@ubc.ca) is an assistant professor in the Department of Political Science, University of British Columbia, Vancouver, BC V6T 1Z1, Canada. Ali Kagalwala (alikagalwala@tamu.edu) is a visiting assistant professor in the Department of Political Science at the Bush School of Government and Public Service, Texas A&M University, College Station, TX 77843. Andrew Q. Philips (Andrew.Philips@colorado.edu) is an associate professor in the Department of Political Science, University of Colorado Boulder, Boulder, CO 80309. Mark Pickup (mark.pickup@sfu.ca) is a professor in the Department of Political Science at Simon Fraser University, Burnaby, BC V5A 1S6, Canada. Guy D. Whitten (g-whitten@tamu.edu) is the Cullen-McFadden Professor of Political Science in the Bush School of Government and Public Service, Texas A&M University, College Station, TX 77843.

Replication files are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the *JOP* replication analyst. An online appendix with supplementary material is available at https://doi.org/10.1086/726934.

Published online December 7, 2023.

The Journal of Politics, volume 86, number 1, January 2024. © 2023 Southern Political Science Association. All rights reserved. Published by The University of Chicago Press for the Southern Political Science Association. https://doi.org/10.1086/726934

provide advice about the benefits of different Monte Carlo performance statistics. There is a seemingly endless combination of such statistics to choose from—such as bias and RMSE or bias and SD. We provide a way to think through what can be learned from various combinations—for example, if an estimator shows no evidence of bias, we explain what might then be gleaned from the SD. Our article also helps the second group, readers of Monte Carlo work, to better understand the trade-offs of various performance statistics and will encourage them to think more critically about the conclusions that can be reached from Monte Carlo simulations. In our literature review, we show that there is tremendous variation in what gets reported. For these readers, we provide useful definitions of the six most common performance statistics. We then offer a structured way to think about what gets reported, what might be missing, and how this should influence our decisions about which estimator or model to use.

To demonstrate the advantages of our recommended approach, we replicate parts of three prominent, recent articles that use Monte Carlo experiments to guide researchers about their choice of methods. In each case, our replication demonstrates that using a broader set of performance statistics provides new insights into the relative merits of rival methods. In two of these instances (Clark and Linzer 2015; Wilkins 2018), we find that the recommended method in the original article may not always be preferred. In the third (Hanmer and Kalkan 2013), although our evaluation of the best performing method remains the same as the one recommended in the original article, we demonstrate that the best performing method is problematic for statistical inference.

We begin with an overview of the use of Monte Carlo experiments in political science and present our argument for when and why researchers should consider different performance statistics when evaluating the relative utility of different methods. We then review the use of performance statistics in papers published in the major political science journals and discuss what is missing. We replicate parts of three prominent articles in political science and conclude with a discussion of how our recommendations should be used in future research.

MONTE CARLO EXPERIMENTS AND PERFORMANCE STATISTICS

Monte Carlo simulations are employed across a broad range of academic and applied disciplines.[1] Political science researchers, like those in other fields (e.g., Hastie, Tibshirani, and Friedman 2009; Robert and Casella 2010), have used Monte Carlo methods for two main purposes—first, for evaluating the performance of rival methods and, second, for the estimation or interpretation of statistical models (e.g., Gill 2014; Jackman 2009). In this article, our focus is on the use of Monte Carlo simulations, also referred to as "Monte Carlo experiments," for the evaluation of the performance of rival methods.

Generically, we can think of Monte Carlo experiments as a staged competition between two or more rival methods of estimating the same quantity of interest, which we will label θ.[2] The standard practice is for θ to be fixed and the data repeatedly simulated from one or more user-created stochastic data-generating processes (DGPs). These DGPs are usually set up to mimic circumstances that applied researchers are likely to encounter. For each sample of data, the rival methods are then used to calculate an estimator, $\hat{\theta}$.[3] Performance statistics are different ways to evaluate the ability of each rival method to accurately reflect the properties of θ across n simulations.

In the remainder of this section, we define and discuss the crucial aspects of the six performance statistics that we recommend for reporting (bias, SD, overconfidence, RMSE, coverage, and power). For each performance statistic, we provide a definition, the relevant formulas (if needed), and a short summary of the statistic's importance.

Bias, standard deviation, and overconfidence

In figure 1, we illustrate bias, SD, and overconfidence for a hypothetical quantity of interest, θ, and estimates, $\hat{\theta}$. We depict the results from a set of hypothetical simulations for an estimator of the true parameter value θ. The bars depict the density of the estimated values of θ and the vertical line in the center of the figure indicates the expected or average value of $\hat{\theta}$.

Bias

Definition and formula. As demonstrated in figure 1, the bias of an estimator for a quantity of interest is defined as the difference between the expected value of the estimates of the quantity from repeated sampling and the value of the quantity

[1] For general overviews of Monte Carlo methods, see Barbu and Zhu (2020) or Thomopoulos (2012).

[2] We refer to θ as a "quantity of interest" to reflect the fact that, while some researchers are focused on the estimation of parameters, others are focused on the performance of test statistics (Philips 2018) or other quantities of interest such as long-run multipliers in time series analyses (Webb, Linn, and Lebo 2020) or indirect effects in spatial analyses (Whitten, Williams, and Wimpy 2021).

[3] Rival methods include different models and estimators. For ease of exposition, we use the term "estimators" from here on so that we do not need to repeatedly write "models and estimators."

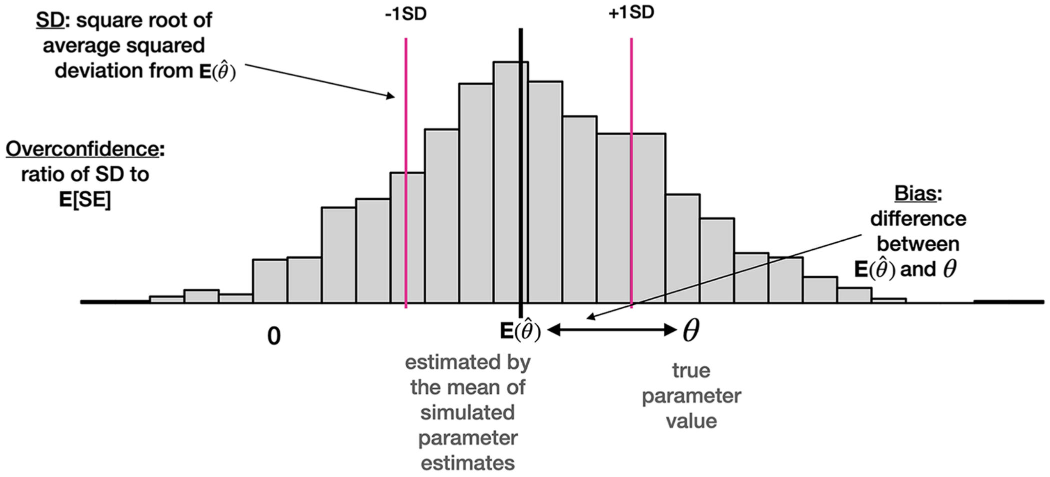
SD: square root of average squared deviation from E($\hat{\theta}$)
-1SD
+1SD
Overconfidence: ratio of SD to E[SE]
Bias: difference between E($\hat{\theta}$) and θ
0
E($\hat{\theta}$)
θ
estimated by the mean of simulated parameter estimates
true parameter value

Figure 1. Bias, SD, and overconfidence

in the DGP. When $E(\hat{\theta}) \neq \theta$, as in the figure, the estimator is biased.

Definition.

$$\text{Bias}[\hat{\theta}] = \text{E}[\hat{\theta} - \theta] = \text{E}[\hat{\theta}] - \theta. \tag{1}$$

Calculation.

$$\widehat{\text{Bias}}[\hat{\theta}] = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta). \tag{2}$$

Bias is typically calculated as the average deviation of the estimates of the quantity of interest from the DGP value. This average is calculated across the simulations. While "average bias" is by far the most commonly calculated quantity, others are possible, including median bias (e.g., Pickup and Hopkins 2022), which is useful when the quantity of interest is not normally distributed (e.g., when calculating nonlinear combinations of parameter estimates for long-run effects in time series). Researchers may also plot the distribution of each estimate's distance from the true DGP value (cf. Helgason [2016], who presents box-whisker plots depicting the distribution of absolute bias from rival estimators in his simulations).

Importance. Calculating bias approximates whether using an estimator in an empirical application would, on average, across applications, produce estimates that are equal to the quantity of interest.[4]

Standard deviation

Definition and formula. The SD of an estimator is the square root of the variance of estimates. An estimator has a smaller variance than another if its dispersion around its expected value is less than that of the other estimator. As depicted in figure 1, this performance statistic measures the square root of the average squared deviation of the values of $\hat{\theta}$ around $\text{E}(\hat{\theta})$.

Definition.

$$\text{SD}[\hat{\theta}] = \sqrt{\text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2]}. \tag{3}$$

Calculation.

$$\widehat{\text{SD}}[\hat{\theta}] = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(\hat{\theta}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i\right)^2\right]}. \tag{4}$$

The variance is calculated as the average squared deviation of the estimates from the average estimate. The SD is calculated as the square root of this value.

Importance. Because researchers usually encounter only one sample from the population, SD informs us how close that quantity is likely to be to $\text{E}[\hat{\theta}]$, which itself may or may not be biased (e.g., $\text{E}[\hat{\theta}]$ may not equal θ). This measure is most useful as a relative comparison between the SD of two or more rival estimators.

Overconfidence

Definition and formula. Overconfidence is used to assess the accuracy of estimated standard errors. As we depict in figure 1, overconfidence is the SD of the estimates divided by

[4] When making relative comparisons of bias across competing estimators, there may not always be an estimator that is unbiased. Thus researchers prefer the estimator that, all else equal, has the lowest bias. For another discussion of the importance of bias, see Carsey and Harden (2014).

the expected value of the estimated standard errors for a quantity of interest.[5]

Definition.

$$\text{Overconfidence}(\hat{\theta}) = \frac{\text{SD}(\hat{\theta})}{\text{E}[\text{SE}(\hat{\theta})]}. \quad (5)$$

Calculation.

$$\widehat{\text{Overconfidence}}(\hat{\theta}) = \frac{\widehat{\text{SD}}[\hat{\theta}]}{\frac{1}{n}\sum_{i=1}^{n} \text{SE}(\hat{\theta}_i)}. \quad (6)$$

Overconfidence is calculated by dividing the calculated SD by the average calculated standard error, across the n simulations. A value of 1 implies accurate standard errors, a value greater than 1 implies overconfidence, and a value less than 1 implies underconfidence.

Importance. Most empirical applications of estimators involve statistically testing a theoretically derived hypothesis against a null hypothesis. In these applications, rejecting the null hypothesis provides evidence in support of the researcher's theory.[6] Overconfidence means that the standard errors are underestimated, which results in smaller confidence intervals that increase the probability of rejecting the null hypothesis when it is true (i.e., we find support for the theory when it is not true). This scenario can also be described as an increase in type 1 errors, which are defined as incorrectly rejecting a true null hypothesis. Underconfidence means that the standard errors are overestimated, which results in larger confidence intervals that decrease the probability of rejecting a false null hypothesis. This scenario can also be described as an increase in type 2 errors, which are defined as incorrectly failing to reject a false null hypothesis.

Root mean squared error, coverage, and power

We illustrate our three other recommended quantities of interest, RMSE, coverage, and power, in figure 2. As in figure 1, we show the density of the estimates of θ with the bars. The dashed line on the left side of figure 2 shows the value of the false null hypothesis, specified as zero, and the dashed line on the right side of figure 2 shows the DGP value of θ.[7] Under the histogram, for eight example estimates ($\hat{\theta}$), we show the point estimate with a 95% confidence interval to illustrate how coverage and power are defined.

Root mean squared error

Definition and formula. RMSE is a measure of the average error of an estimator.[8] As shown in figure 2, it is defined as the square root of the expected value of the squared differences between the estimates and the true value. Alternatively, it can be expressed as the square root of the sum of squared bias and the variance of an estimator. RMSE is the combination of bias and SD, so lower values of RMSE are preferred.

Definition.

$$\text{RMSE}[\hat{\theta}] = \sqrt{\text{E}[(\hat{\theta} - \theta)^2]} = \sqrt{\text{Bias}(\hat{\theta})^2 + \text{SD}^2(\hat{\theta})}. \quad (7)$$

Calculation.

$$\widehat{\text{RMSE}}[\hat{\theta}] = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(\hat{\theta}_i - \theta)^2]}. \quad (8)$$

RMSE is calculated by taking the square root of the average squared difference between the estimates and the true value.

Importance. As is the case with SD, RMSE is most useful for relative comparisons between two or more estimators. When evaluating the performance of rival estimators, researchers may find themselves with estimators that vary in terms of bias and variance and thus face a bias-variance trade-off. For example, in the presence of unobserved time-invariant unit heterogeneity that is correlated with the regressors, the fixed effects estimator is unbiased but has a larger SD, and the random effects estimator is biased but has a smaller SD (Clark and Linzer 2015). As a result, researchers may use RMSE to evaluate whether the losses in accuracy from one estimator are larger than those from other estimators.[9]

[5] Researchers may alternatively calculate standard error bias, which is defined as $\text{E}[\text{SE}(\hat{\theta})] - \text{SD}(\hat{\theta})$. This would be used in the same situations as eq. (5). See the appendix for a discussion on the relationship between our measure of overconfidence and others in the literature (e.g., Beck and Katz 1995; Franzese and Hays 2007).

[6] Other empirical applications include theories that predict a null result. In such cases, failing to reject the null hypothesis provides evidence for the researcher's theory. See Rainey (2014) for an explanation on how researchers can evaluate theories that predict a null effect.

[7] As discussed in the appendix, power is dependent on the specification of the null hypothesis, most commonly 0 as shown in fig. 2.

[8] It is noteworthy that RMSE is only one possible weighted combination of bias and variance. Researchers may choose other weighted combinations of bias and variance based on their requirements.

[9] Since RMSE is a function of both bias and SD, it may seem redundant that we recommend researchers calculate all three performance statistics. See below for a discussion of why calculating all three performance statistics is important.

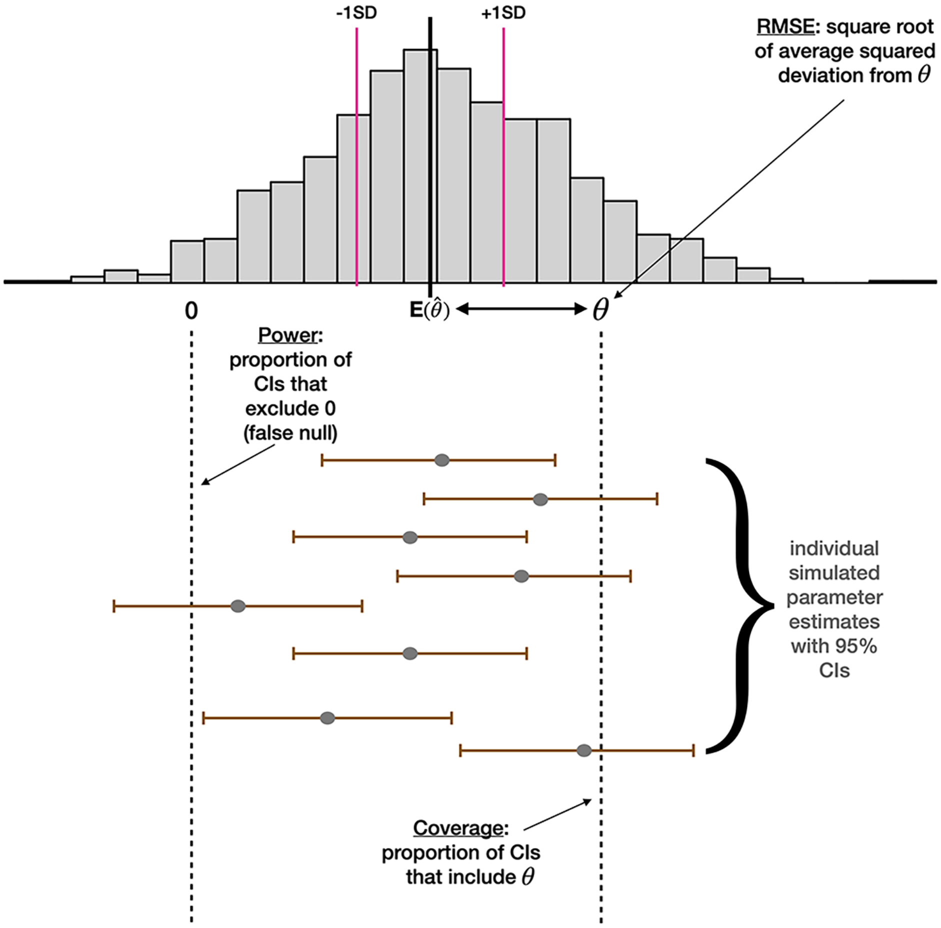
-1SD
+1SD
RMSE: square root of average squared deviation from θ
0
$E(\hat{\theta})$
θ
Power: proportion of CIs that exclude 0 (false null)
individual simulated parameter estimates with 95% CIs
Coverage: proportion of CIs that include θ

Figure 2. RMSE, coverage, and power

Coverage probability

Definition. As we illustrate in figure 2, coverage probability is the proportion of times the confidence intervals of the estimator encompass the true DGP value. It is calculated as the proportion of estimated confidence intervals that contain the DGP value. If the eight confidence intervals depicted in figure 2 were the only simulations that had been carried out, the coverage probability would be 0.375 since only three of the depicted intervals include the dashed line for θ. In practice, researchers typically would conduct many more than eight simulations and thus have many more than eight confidence intervals. If the 95% confidence interval is correctly sized, we expect that in a large number of repeated samples, the constructed 95% confidence intervals will not overlap with the true effect 5% of the time (Jackman 2009).[10] Thus, one should expect a coverage probability of 0.95 if one is using 95% confidence intervals. Coverage probabilities larger than 0.95 mean that the estimated confidence intervals encompass the true null hypothesis more often than expected, while coverage probabilities less than 0.95 mean that the estimated confidence intervals encompass the true null hypothesis less often than expected.

Importance. High (low) coverage probability means a lower (higher) type 1 error rate ($\Pr(\text{type 1 error}) = 1 - \text{coverage}$). However, higher coverage probability is not always better.[11] Researchers should prefer coverage probabilities closer to the confidence level (e.g., a 0.95 coverage probability for the 95% confidence level). Coverage probability informs researchers about the probability that an estimator will reject the true null hypothesis and incorrectly conclude in favor of the alternative hypothesis (type 1 error).

Power

Definition. The power of an estimator is the proportion of instances in which the null hypothesis is correctly rejected.

[10] In the appendix we provide some further details on the relationship between coverage probability, power, and relevant researcher choices of hypothesis test specification.

[11] For example, a coverage probability greater than 0.95 at the 95% confidence level indicates overestimated standard errors.

In other words, as we depict in figure 2, power is the proportion of instances in which the confidence intervals reject the false null hypothesis. It is calculated as the proportion of estimated confidence intervals that do not contain the null hypothesis. If the eight confidence intervals depicted in figure 2 were the only simulations that had been carried out, the power would be 0.875 since only one of the eight confidence intervals includes the dashed line for 0, the false null hypothesis value in this hypothetical illustration. As we noted in our discussion of coverage probability, researchers typically would conduct many more than eight simulations and thus have many more than eight confidence intervals.

Importance. Low power translates into a high incidence of type 2 errors ($\Pr(\text{type 2 error}) = 1 - \text{power}$). Failing to reject the null hypothesis when it is false results in incorrect inferences about the plausibility of the alternative hypothesis. As a result, all else equal, it is important that an estimator has high power. While coverage probability informs us whether we can be confident that an estimator will not incorrectly reject the null hypothesis when it is true, power informs us as to whether the estimator will correctly reject the null hypothesis when it is false.

APPLYING THE PERFORMANCE STATISTICS

The value of the six performance statistics that we defined in the previous section will vary across applications. Nonetheless, it is useful to think about the value of the performance statistics that we recommend in general terms and, in particular, to think about the value of the different performance statistics in combination with each other. To do this, we divide our recommended performance statistics into two groups of three.

The first group of performance statistics—RMSE, coverage probability, and power—*evaluates* an estimator's performance on point estimates and inference. The second group of performance statistics—bias, SD, and overconfidence—helps to *diagnose* why an estimator has large or small average error (RMSE), why it has high or low coverage probability, and why it has high or low power. We recommend that researchers begin by using the first group of performance statistics to evaluate how an estimator performs in terms of point estimates and inference and then, if needed, diagnose and understand these results using the second group of performance statistics.[12]

Evaluate

As depicted in figure 3, we divide the evaluation of estimator performance into point estimates and inference. To be clear, we expect most producers and readers of Monte Carlo experiments to be interested in both the point estimate and inference performances of estimators.

Point estimates. RMSE is a summary measure of how much point estimates differ from the true DGP value because of the systematic over- or underestimation of an estimator (bias) and the sampling variability (SD). It thus summarizes overall how far off the estimate will be, on average, from the true value. This is valuable information when comparing the strengths and weaknesses of different estimators for point estimates.

Inference. Coverage probability and power inform researchers whether type 1 and type 2 errors will be inflated, respectively. These are both important pieces of information when comparing the strength and weaknesses of different estimators for hypothesis-testing inferences.

Diagnose

The second step in figure 3 is to diagnose the sources of interesting performances from our evaluation step. While RMSE, coverage probability, and power provide useful summaries of how well the estimator will perform with respect to point estimates and hypothesis-testing inferences, they obscure exactly why an estimator performs well or poorly. This is because they are each a function of multiple fundamental properties of the estimator. Below, we describe how bias, SD, and overconfidence help diagnose poor performance with respect to RMSE, coverage probability, and power.

RMSE. If the RMSE is small, this tells us the bias and SD are small.[13] However, if the RMSE is not small, it does not reveal whether this is caused by large bias, large SD, or both. It is also possible that two estimators will have a similar RMSE even if their bias and SD are substantially different; again, whether bias and SD differ across estimators is hard to know without directly calculating these two performance measures.

[12] This does not necessarily mean starting with RMSE. For example, if a study compares the performance of different robust standard errors and we know that all of the estimators under consideration are unbiased, then we do not recommend starting with RMSE. We do note that coverage and power are important statistics to understand the performance of such robust standard errors. We also note that if there is poor coverage or power, then overconfidence (and SD) can shed light on why this is the case. Yet, if simulations show no problems with power and coverage (or they are good enough that we are comfortable with the performance of the robust standard error), then we can be confident there are no problems with overconfidence.

[13] By "small," we generally mean close enough to zero that we expect estimates to be within the precision of our original measures.

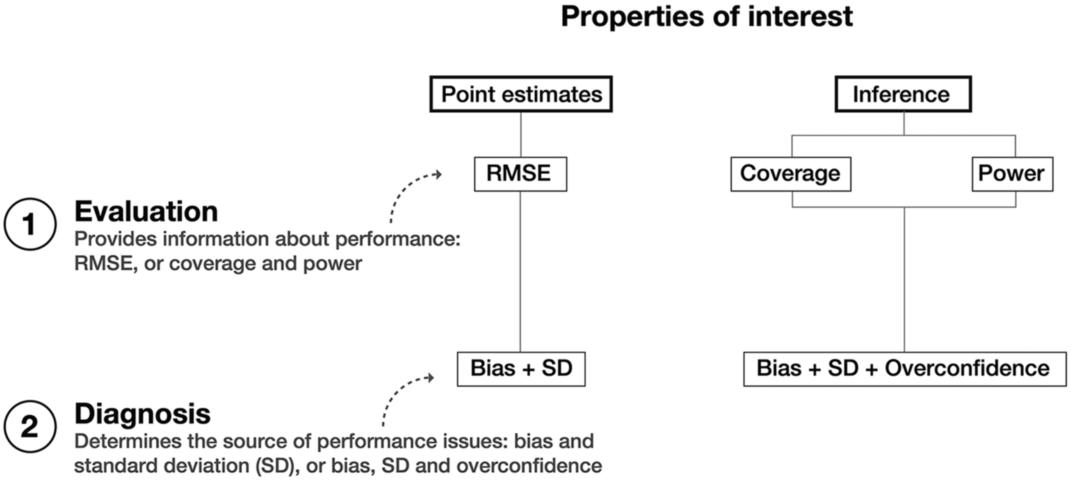
Properties of interest
Point estimates
Inference
RMSE
Coverage
Power
1
Evaluation
Provides information about performance: RMSE, or coverage and power
Bias + SD
Bias + SD + Overconfidence
2
Diagnosis
Determines the source of performance issues: bias and standard deviation (SD), or bias, SD and overconfidence

Figure 3. Information provided by performance statistics

Examining bias is valuable because it tells us on average how well an estimator will perform. A large bias means that an estimator will perform poorly even if the researcher has taken steps to minimize random error, for example, with a large sample size. However, this has limitations. Even if the estimates from repeated sampling are equal to the true value in the DGP on average, this does not imply that the quantity estimated from one sample is going to be equal to or close to the true parameter value. In reality, researchers usually encounter only one sample drawn from the underlying population. Fortunately, the SD informs us whether the estimated quantity of interest from a given sample is likely to be closer to or farther away from the average estimate, although it cannot tell us whether this average estimate will be close to the true value. Therefore, in order to diagnose the source of large RMSE in the point estimate of an estimator, both bias and SD need to be examined in combination.[14]

Coverage probability and power. The location and width of confidence intervals are a function of bias and standard errors, the latter of which are estimates of SD. As such, both power and coverage probability are determined by bias, SD, and overconfidence or some combination of the three. Note though that SD is probably the least valuable of these three statistics when considering coverage probability. If there is no bias, the degree of SD will have no effect on coverage probability, except to the extent that it affects overconfidence; underestimated standard errors will result in a lower coverage probability. Consider another scenario in which there is bias. A larger SD might mitigate the effects of bias but only inadvertently. For example, if your estimate is very far off from the true parameter value, the confidence interval may still include the true parameter if there is a great degree of reported uncertainty in your estimate. In other words, the SD will not be a source of poor coverage probability, but it might explain why a badly biased estimator may still have a good coverage probability. With respect to power, smaller SD or overconfidence should increase power, but the latter does so by incorrectly estimating the precision of the estimate. Holding all else constant, attenuation bias ($0 < |\mathrm{E}[\hat{\theta}]| < |\theta|$) will lower power. Consider a scenario in which there is attenuation bias, high SD, and underconfidence. In this case, power will be less in contrast to when bias is absent. Inflationary bias ($0 < |\theta| < |\mathrm{E}[\hat{\theta}]|$) will increase power but at the expense of a poor estimate, on average. Overall, in order to diagnose the source of problems of inference due to poor power or coverage probability, we recommend examining bias, overconfidence, and SD in combination.

Choosing which performance statistics to report

Given the value of the measures for evaluating and diagnosing the performances of rival estimators, we recommend the reporting of all six. We recognize, however, that journal space is limited and that some authors and journal editors are inclined to hold the line on the increasingly large supplemental materials documents that accompany published papers. With this in mind, we provide a guide on which performance statistics to report:

1. Evaluate the estimators on RMSE, coverage probability, and power. Use this to identify estimators

[14] It is true that bias can be calculated from SD and RMSE, and SD can be calculated from bias and RMSE, but this involves a substantial effort on behalf of readers. Further, because RMSE is a nonlinear combination of SD and bias, it is only by reporting both SD and bias that the relative contribution of each to RMSE is clear.

that perform poorly and differently with respect to point estimates (RMSE) or inference (coverage probability or power). If the estimators perform well or similarly on one or more performance statistics, those results need only a brief mention.

2. Diagnose the estimators that perform poorly or differently on RMSE, coverage probability, and power, using the appropriate combination(s) of bias, SD, and overconfidence, as per figure 3. If the estimators perform well or similarly, we recommend a brief summary of these results. Otherwise, if the estimators perform poorly and differently across these diagnostic performance statistics, then we recommend that researchers present the results of these diagnostics in more detail.

We recognize that oftentimes the above guide will lead to the reporting of all six performance statistics. However, this is not always the case. For example, consider Philips (2022), who generates two independent unit roots in one of his Monte Carlo experiments and compares the performance of three time series models in terms of type 1 error rates of the long-run effects—lagged dependent variable model (LDV), error correction model (ECM), and autoregressive distributed lag model, or ADL(1,1), with one lag of the dependent variable and the regressor.[15] He finds that all three models perform similarly, and poorly, in terms of the coverage probability of the long-run effect. Following our recommendations, he should summarize the results for coverage probability briefly in text, for example, "I find that all three models perform similarly with a rejection rate of around 0.2," and then present the diagnostic performance statistics—bias, SD, and overconfidence—that result in such type 1 error findings using figures or tables.

When presented with a marginal choice between reporting all six performance statistics and saving journal/appendix space, we believe that, in an era in which replication files and online appendixes are the norm, the cost of reporting all six performance statistics is outweighed by the benefit of providing a more comprehensive understanding of an estimator to readers. As we demonstrate later in this article, when examining all six performance statistics, we can learn novel and important things about estimators that may lead to conclusions about the preferred estimator different from those of the original author(s). Before turning to these replications, we review current practices and what is missing.

PATTERNS OF REPORTING PERFORMANCE STATISTICS

In order to assess the degree to which our recommended performance statistics are currently being used by political science researchers in their Monte Carlo simulations, we had two research assistants each code every published article in the *American Journal of Political Science*, the *American Political Science Review*, the *Journal of Politics*, *Political Analysis*, and *Political Science Research and Methods* from 2006 to 2016 that contained the keywords "Monte Carlo" or "simulation."[16]

To get a sense of which performance statistics are being reported and how they are being reported together, we present the most common patterns of reporting for our recommended performance statistics in table 1.[17] Each row in table 1 depicts a different combination of performance statistic reporting that we found in our coding, listed in order from the most to least common. As we can see from this table, the modal pattern was to report only bias, while the second most popular pattern was to report both bias and RMSE. Looking at the bottom row of table 1, we can see that in terms of overall use, bias was by far the most reported performance statistic, being present in 85.9% of the studies, followed by RMSE or mean squared error (MSE), coverage probability/type 1 error rate, SD, overconfidence, and power/type 2 error rate.

In the far right column of table 1, we provide a short summary of what is missing or unknown when researchers use each pattern of reporting based on our discussion in the previous section. Note how adding bias or SD to RMSE provides additional information. Adding each independently tells us about how one or the other contributes to RMSE, but adding both bias and SD to RMSE gives a much more complete picture of the sources of RMSE. Because coverage probability and power are nonlinear combinations of bias, SD, and overconfidence, it is even more important to provide all three determinants of coverage probability and power to understand the sources of these important inferential properties. Last, we also recognize that tables are not the only way to report Monte Carlo results; some researchers (e.g., Esarey 2016; Helgason 2016; Honaker, Katz, and King 2002) visually show more than one quantity of interest—for instance, bias

[15] LDV: $y_t = \alpha + \phi y_{t-1} + \beta_1 x_t + \varepsilon_t$, ECM: $\Delta y_t = \alpha + \phi y_{t-1} + \beta_1 \Delta x_t + \beta_2 x_{t-1} + \varepsilon_t$, and ADL(1,1): $y_t = \alpha + \phi y_{t-1} + \beta_1 x_t + \beta_2 x_{t-1} + \varepsilon_t$.

[16] Since publication of *Political Science Research and Methods* began in 2013, we coded 2013–16 for that journal. We coded all Monte Carlo simulations that were presented as a part of published papers and in appendixes that appeared as a part of the volume in which they were published; see the appendix for details.

[17] See the appendix for the full table and additional details. We also found a very small number of papers that reported performance statistics other than those listed in table 1.

Table 1. Most Common Patterns of Reporting Performance Statistics in Major Political Science Journals

Performance Statistic							
Bias	RMSE or MSE	Coverage/ Type 1	SD	Overconfidence	Power/ Type 2	Pattern (%)	Missing (Unknown)
B						12.7	Average error and one source; inference problems and two sources
B	R					9.9	One source of average error; inference problems and two sources
	R					8.5	Sources of average error; inference problems and their sources
B		C				7.0	Average error and one source; power; two sources of inference problems
B			S			5.6	Average error; inference problems and one source
B		C			P	5.6	Average error and one source; two sources of inference problems
B		C	S			5.6	Average error; power and one source of inference problems
B	R			O		5.6	One source of average error; inference problems and one source
B				O		4.2	Average error and one source; inference problems and one source
B			S	O		4.2	Average error; inference problems
B	R	C				4.2	One source of average error; power; two sources of inference problems
85.9	45.1	36.6	29.6	23.9	19.7	Overall use (%)	

Note. Letters indicate the particular performance statistic was reported for studies referenced in that row. B = bias, R = RMSE, C = coverage probability, S = SD, O = overconfidence, P = power. See the appendix for the full table of reporting patterns.

as well as percentiles of the estimates and outliers—through the use of box-whisker plots.

THREE REPLICATIONS

As we demonstrated in the previous section, political science researchers usually use three or fewer performance statistics in their Monte Carlo experiments. While table 1 provides a brief summary of what is missing or unknown with each of the observed patterns, in this section we take a closer look by using the diagram presented in figure 3 to replicate and extend the analyses of three prominent articles that use Monte Carlo simulations to assess the relative utility of different estimators. In each case, the use of additional performance statistics would have changed, refined, or more strongly supported their conclusions regarding the desirability of different estimators. We first replicate Clark and Linzer (2015) and provide a full example of following our recommendations. Our second and third replications are of Wilkins (2018) and Hanmer and Kalkan (2013), respectively. We report only a summary of these findings and provide full details in the appendix.

Clark and Linzer replication

Clark and Linzer (2015) weigh in on the debate between using unit intercepts (i.e., fixed effects) or random unit intercepts (random effects) to address the issue of time-invariant unobservable individual effects in panel data. As the authors state, random effects tend to have a lower variance than fixed effects, but with the strong assumption that "the random-effects estimator requires there to be no correlation between the covariate of interest, x, and the unit effects" (402). Clark and Linzer use RMSE as a measure of estimator performance across a range of values for J (number of units) and n (number of within-unit observations) common in the social sciences using the following DGP:

$$y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_y^2), \beta = 1, \tag{9}$$

$$x_i \sim N(\bar{x}_j, \sigma_x^2), \tag{10}$$

$$\begin{bmatrix} \alpha_j \\ \bar{x}_j \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \tag{11}$$

where α_j are the unit intercepts and x_i are within-unit values of the independent variable drawn from a normal distribution with unit mean $\bar{x}_j$ and variance σ_x^2. The independent and identically distributed error term is ε_i, with mean 0 and variance σ_y^2. The within-unit means $\bar{x}_j$, and unit intercepts α_j, are drawn from a multivariate normal (MVN) distribution with mean 0, variance 1, and covariance between $\bar{x}_j$ and α_j equal to ρ ($\rho = 0, 0.1, 0.2, \ldots, 0.9, 0.95$). Across these conditions, they compare the relative performances for the following three models: feasible generalized least squares random effects (FGLS-RE), ordinary least squares with fixed effects (OLS-FE), and ordinary least squares with no adjustments for the nature of the data (OLS-pooled).

Clark and Linzer find that when within-unit variation is small ($\sigma_y = 1$ and $\sigma_x = 0.2$ in their simulations), the number of within-unit observations (n) is small, and the amount of correlation between the unit intercepts and the independent variable (ρ) is low, the RMSE of the FGLS-RE estimator is lower than that of the OLS-FE estimator. That is to say, even though the assumption underlying random effects has been violated, the gain in efficiency still outweighs the increase in bias. The authors thus conclude that we should prefer random effects over fixed effects under these conditions. However, as ρ increases, the random effects estimator performs much worse than the fixed effects estimator in terms of RMSE.

While using RMSE is a good way to examine both bias and efficiency in a single statistic, we argue that using a single statistic to evaluate performance between estimators is at best somewhat limited and at worst potentially misleading as to the best model under particular circumstances. To demonstrate this, we replicate Clark and Linzer's "sluggish" Monte Carlo example, in which x has low within-unit variance ($\sigma_x^2 = 0.2$).[18] Using the variables (α_j and $\bar{x}_j$) from equation (11), we then generated the dependent variable y for unit j at a given within-unit observation i, from equations (9) and (10). Following the procedure of the authors, we simulated 2,000 data sets across values of ρ, while $J = 10, 40, 100$ and $n = 5, 20, 50$ were varied. We then estimate OLS-pooled, OLS-FE, and FGLS-RE models.

In accordance with our recommendations in figure 3, we begin by evaluating the estimators using RMSE, coverage probability, and power. Figure 4 shows the RMSE results for $\hat{\beta}$ from the simulations. These results are identical to figure 2 in Clark and Linzer (2015, 406). As is clear from figure 4, the OLS-FE estimator (*solid line*) is able to produce an RMSE that remains constant as ρ—the correlation between the unit intercepts and $\bar{x}_j$—varies. In contrast, higher levels of ρ tend to increase the RMSE for both the FGLS-RE (*dashed line*) and the OLS-pooled (*dotted line*) estimators. Despite this, when $n = 5$, both the pooled and random effects models tend to outperform the fixed effects estimator when ρ is low. The same holds for random effects, but not the pooled model, when $J = 10$; if ρ is low enough, the random effects estimator performs as well, or better than, the fixed effects estimator. It is only when J and n become large (n in particular) that the fixed effects estimator always outperforms the other two estimators. Thus, were we to only rely on figure 4, we would reach the same conclusions as Clark and Linzer, namely, that when within-unit variation in x is small, there are conditions under which the random effects estimator may be preferred to fixed effects, even when the assumptions underlying the former are violated (when ρ is small but not zero) and n is small.

In figure 5, we show the coverage probability statistics of the estimators (i.e., how often the 95% confidence intervals include the DGP value of $\beta = 1$). Across all levels of ρ, the coverage probability of the fixed effects estimator remains constant at .95. Coverage for the random effects estimator is only that high when $\rho = 0$. When the correlation between the unit effects and the independent variable is nonzero, the random effects model has a lower coverage probability (increased type 1 error); in fact, at high levels of ρ, the coverage probability of the random effects estimator approaches zero. It should also be noted that, across the board, the pooled model performs worse on coverage probability than the fixed effects estimator and worse or as bad as the random effects estimator.

In figure 6 we consider the power of the estimators; that is, how often do they (correctly) reject the false null that $\beta = 0$? For the most part, all three estimators have enough power to reject the null hypothesis when J is greater than 40 and n is greater than 20. However, when n and J are small, the power of the fixed effects estimator is substantially lower than that of the random effects or pooled estimators. This means that the fixed effects estimator will often fail to reject the false null hypothesis (increased type 2 error) when presented with smaller samples, thus leading to incorrect hypothesis-testing inferences.

As noted in figure 3, RMSE, coverage probability, and power merely evaluate estimators' performances with respect to point estimates and inferences and do not provide any information about the reasons behind such performance. To diagnose such performance, we recommend that researchers calculate bias and SD to diagnose sources of the average error in the model and bias, SD, and overconfidence

[18] We also calculate our recommended performance statistics when $\sigma_x = 1$, what Clark and Linzer call the standard case, in the appendix. Our overall conclusions remain the same.

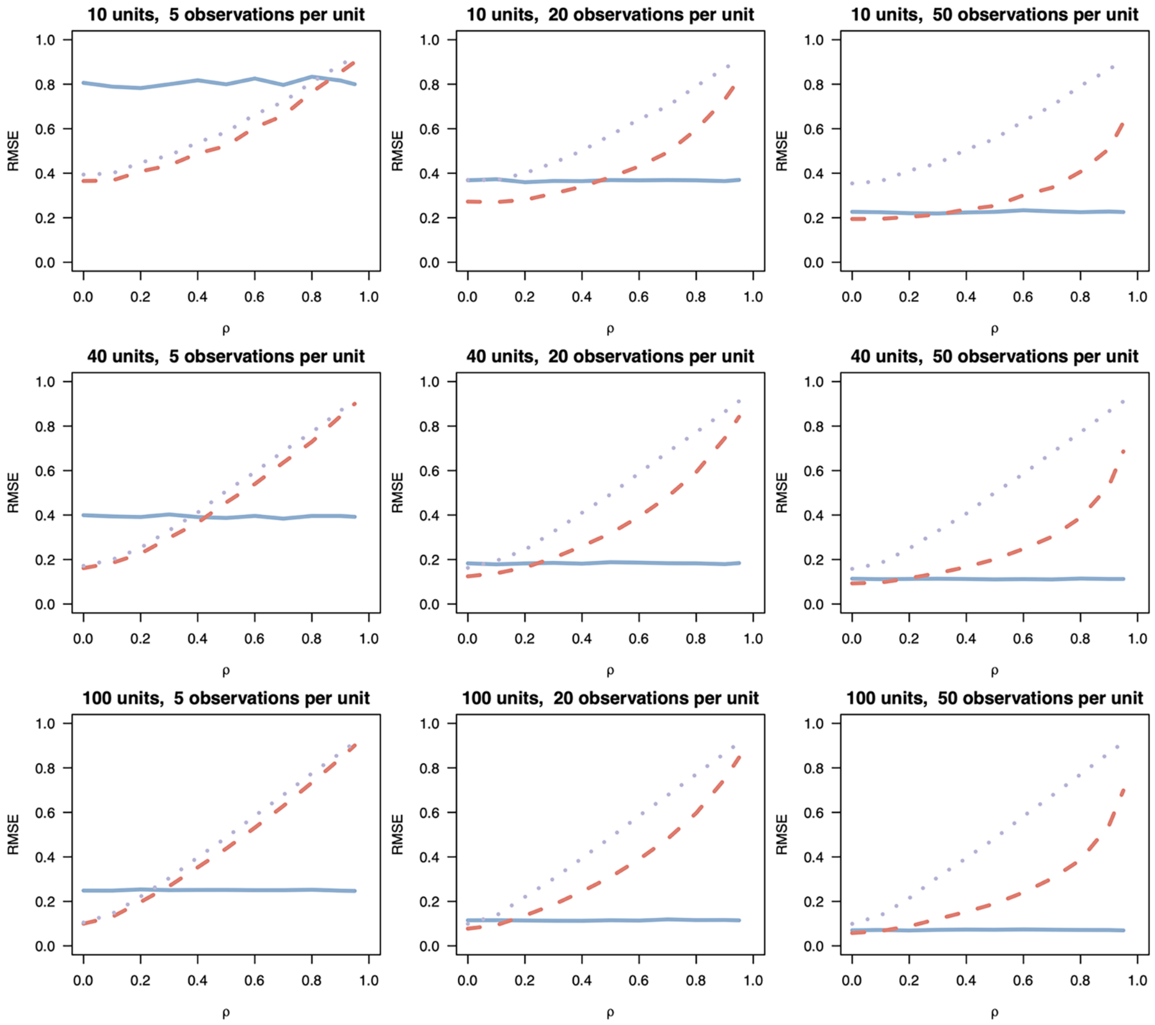

Figure 4. RMSE of $\hat{\beta}$, Clark and Linzer's sluggish case. Solid line = OLS-FE, dashed line = FGLS-RE, dotted line = OLS-pooled model; the horizontal axis is the value of correlation between $\bar{x}_j$ and α_j (ρ).

to diagnose sources of poor coverage probability and power. In figure 7, we show the bias of each estimator for the same simulations. These results demonstrate that, as expected, at any level of correlation between $\bar{x}_j$ and α_j (ρ), the fixed effects estimator is either very slightly biased or unbiased and performs similarly to or better than the pooled and random effects estimators. Together with the results in figure 4, this implies that the fixed effects estimator's RMSE is largely influenced by SD. The pooled and random effects estimators are always biased for any nonzero value of ρ, and this bias increases as the value of ρ increases. Only when $\rho = 0$ and there are 10 unit and five within-unit observations do the pooled and random effects estimators perform better than fixed effects, and only by a small amount. For any nonzero ρ, random effects always performs better than the pooled model. Overall, in terms of bias, the fixed effects estimator performs best when taking into consideration the range of values of J, n, and ρ selected by Clark and Linzer.

Figure 8 shows the SDs from the three models. From this figure, we can see that the efficiency gains from the random effects estimator are greatest when n and J are very small (top-left panel in fig. 8). These relative gains in efficiency decrease as both J and n increase, and the SDs of the fixed effects and random effects estimators are very similar at $n = 50$. The pooled estimator almost always has a lower SD than the fixed effects estimator for $n < 50$, and the SD of the pooled estimator converges to that of the random effects estimator as both J and ρ increase. When comparing figures 7 and 8 we find support for Clark and Linzer's theoretical claim that, under certain conditions, the efficiency gains from the pooled and random effects estimators outweigh their increased bias to produce RMSEs that are lower than those of the fixed effects estimator.

Following our advice in figure 3, in order to determine why the fixed effects estimator performs well and the pooled and random effects estimators perform poorly in terms of

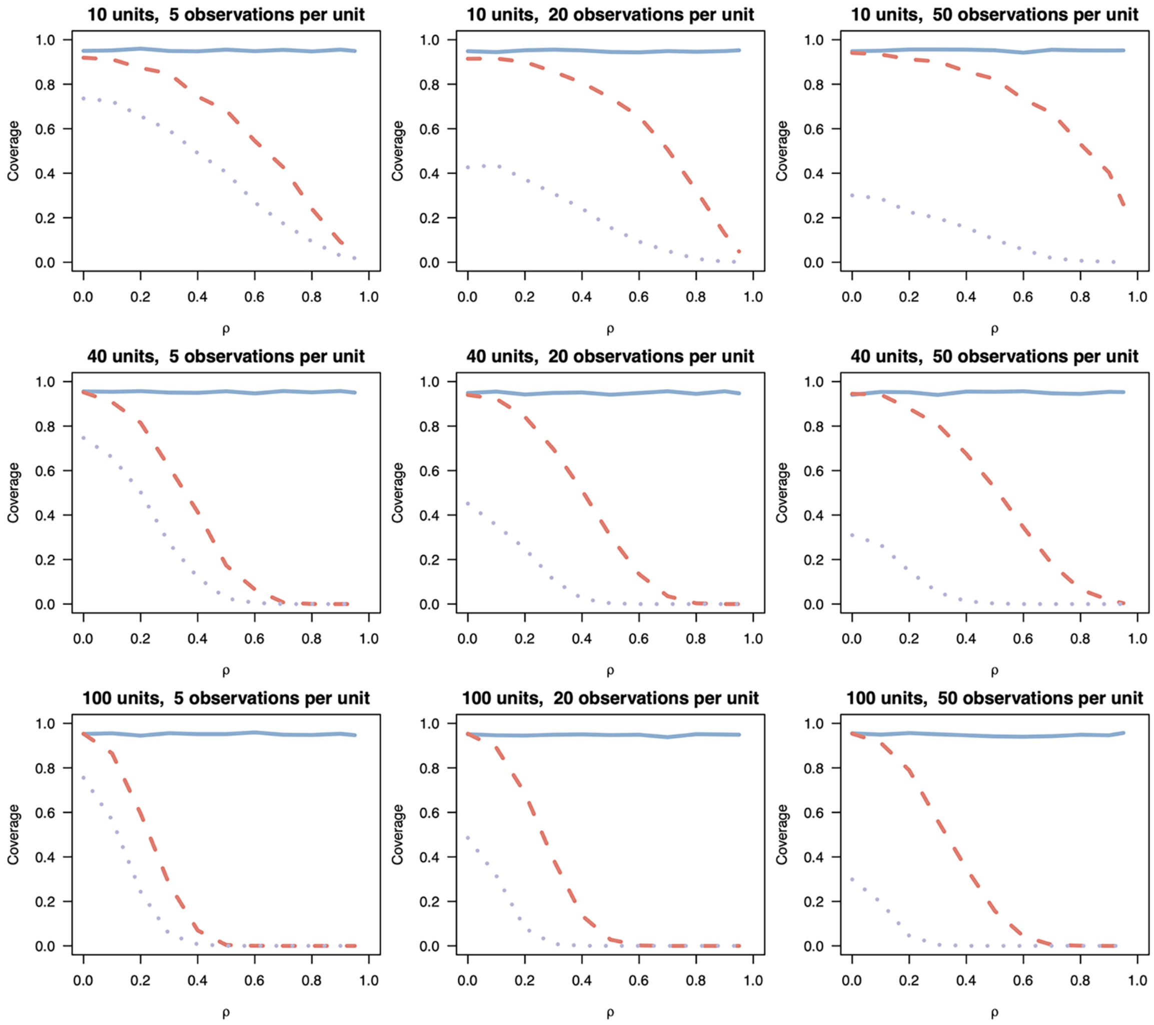

Figure 5. Coverage probability of $\hat{\beta}$, Clark and Linzer's sluggish case. Solid line = OLS-FE, dashed line = FGLS-RE, dotted line = OLS-pooled model; the horizontal axis is the value of correlation between $\bar{x}_j$ and α_j (ρ).

coverage probability, we must analyze overconfidence in addition to bias and SD. Figure 9 demonstrates whether the estimators' standard errors are accurate. As we discussed above, this is an assessment of whether the overconfidence measure differs from 1. From figure 9 we can see that across the board, the pooled estimator is overconfident. When combined with the bias that we see in figure 7, this overconfidence in the pooled estimator results in smaller confidence intervals that are less likely to encompass the true β, resulting in poor coverage probability and increased type 1 errors. When $n < 20$, we can see that the poor coverage probability of the random effects estimator is mainly a function of bias. However, when $n \geq 20$ and $\rho > 0.4$, the random effects estimator's poor coverage probability is a result of both its bias and overconfidence. The random effects estimator only recovers accurate estimates of the SD when $J > 10$ and $n = 5$ or at low levels of ρ when $J > 10$ and $n > 5$. These results combined with the random effects estimator's low bias at low values of ρ result in a high coverage probability at these values of ρ. And, when $J > 10$ and $n = 5$ across high levels of ρ, poor coverage probability is largely a result of increasing bias. The fixed effects estimator, overall, always recovers accurate estimates of the SD of the sampling distribution. Thus, even though the fixed effects estimator has a relatively larger SD, its good coverage probability occurs because it is unbiased and recovers accurate estimates of the SD of the sampling distribution.

By comparing figures 6 and 8, we can see that the panels in which the fixed effects estimator has low power are also the panels in which the fixed effects estimator has large SD values.[19] And, since we know from figure 9 that the fixed effects estimator recovers accurate standard errors across the board, the fixed effects estimator has large confidence intervals due to

[19] These large SD values are due to the constrained variance analyzed by this estimator that only leverages within-unit variation.

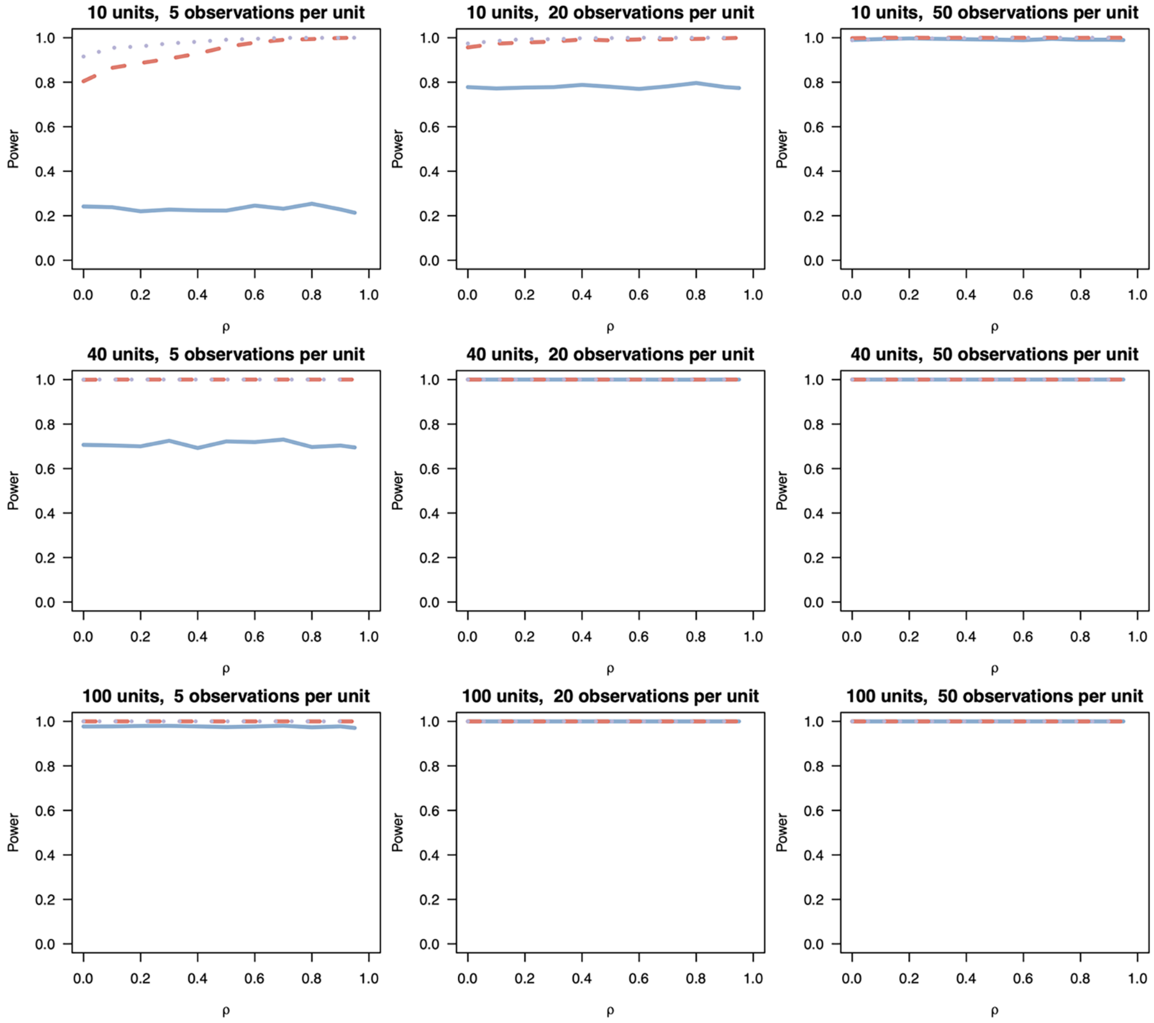

Figure 6. Power of $\hat{\beta}$, Clark and Linzer's sluggish case. Solid line = OLS-FE, dashed line = FGLS-RE, dotted line = OLS-pooled model; the horizontal axis is the value of correlation between $\bar{x}_j$ and α_j (ρ).

its SD, making it more likely that the estimator encompasses the parameter value specified in the false null hypothesis. For the pooled and random effects estimators, their bias, low SD values, and underestimated standard errors result in smaller confidence intervals that are unlikely to encompass 0, the false null hypothesis. This results in high power. When $J = 10$ and $n = 5$, the pooled and random effects estimators have power less than 1. This is because at $\rho = 0$, when both the pooled and random effects estimators are unbiased, the lower power is likely to be due to the small sample size.[20] And, as ρ increases, these sample size issues are masked by increasing bias that moves estimates away from zero, making the rejection of the false null hypothesis more likely.

There are several conclusions to draw from our replication and extension of Clark and Linzer's findings to include measures of bias, SD, power, coverage probability, and overconfidence. First, we are able to exactly replicate their analyses of RMSE. Second, from an analysis of bias and SD, in line with Clark and Linzer's theoretical expectations, we find that the fixed effects estimator's relative inefficiency contributes to its RMSE and that the bias of the pooled and random effects estimators makes a relatively larger contribution to their RMSE values. Third, the good coverage probability of the fixed effects estimator is because of its unbiasedness and ability to recover accurate standard errors, despite having a relatively larger SD. The poor coverage probability of the pooled and random effects estimators are because of their bias and overconfidence. Last, all three estimators perform well on power. The main exception to this is for the fixed effects estimator at low values of J and n. It is worth noting, however, that sometimes the random effects and pooled models perform well on power only because of their sizable bias.

[20] See the appendix for an in-depth discussion of how power is determined.

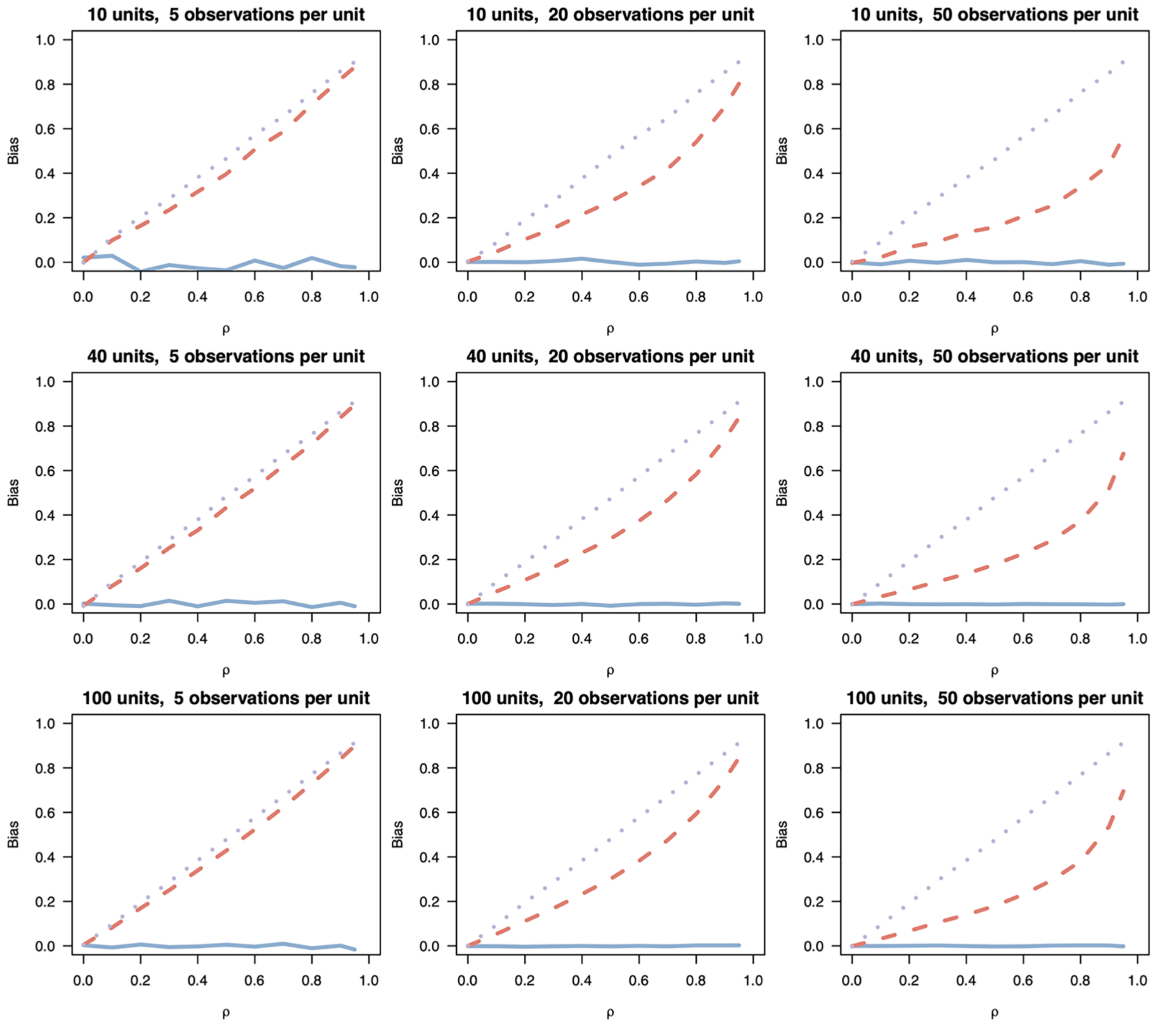

Figure 7. Bias of $\hat{\beta}$, Clark and Linzer's sluggish case. Solid line = OLS-FE, dashed line = FGLS-RE, dotted line = OLS-pooled model; the horizontal axis is the value of correlation between $\bar{x}_j$ and α_j (ρ).

Clark and Linzer (2015, 407) write in their conclusion that "examining the RMSE of both estimators, however, we demonstrate that there is a range of conditions under which it may be worth accepting the bias in the random-effects model if it is associated with a sufficient gain in efficiency, leading to estimates that are closer, on average, to the true value in any particular sample." While we agree with this conclusion in terms of considerations of point estimates only, most researchers are also interested in hypothesis-testing inferences. When we diagnose performances on inference, we reach dramatically different conclusions. This is the case because we find that the fixed effects estimator substantially outperforms its rivals on coverage probability. To prefer the random effects estimator, an applied researcher interested in inference would have to have a small number of observations per unit and put a very high premium on type 2 error (power) over type 1 error (coverage probability) and SD over bias or be extremely confident that $\rho = 0$ (although, outside of simulated data scenarios, ρ is unknowable).[21]

Summary of replications of Wilkins (2018) and Hanmer and Kalkan (2013)

In this section, we provide a brief overview of what we found in our replications and extensions of Wilkins (2018) and Hanmer and Kalkan (2013). In our appendix we provide a full discussion and results from these two replications.

Using a DGP in which the autoregression in the error term varies between 0 and 0.5, Wilkins (2018) compares the percentage bias and average error (RMSE) in the short-run

[21] It is worth noting, however, that from a time series perspective Clark and Linzer's DGPs are all static. A recent article by Plümper and Troeger (2019) demonstrates that some fixed effects estimators can lead to substantial problems if the underlying dynamics have been misspecified.

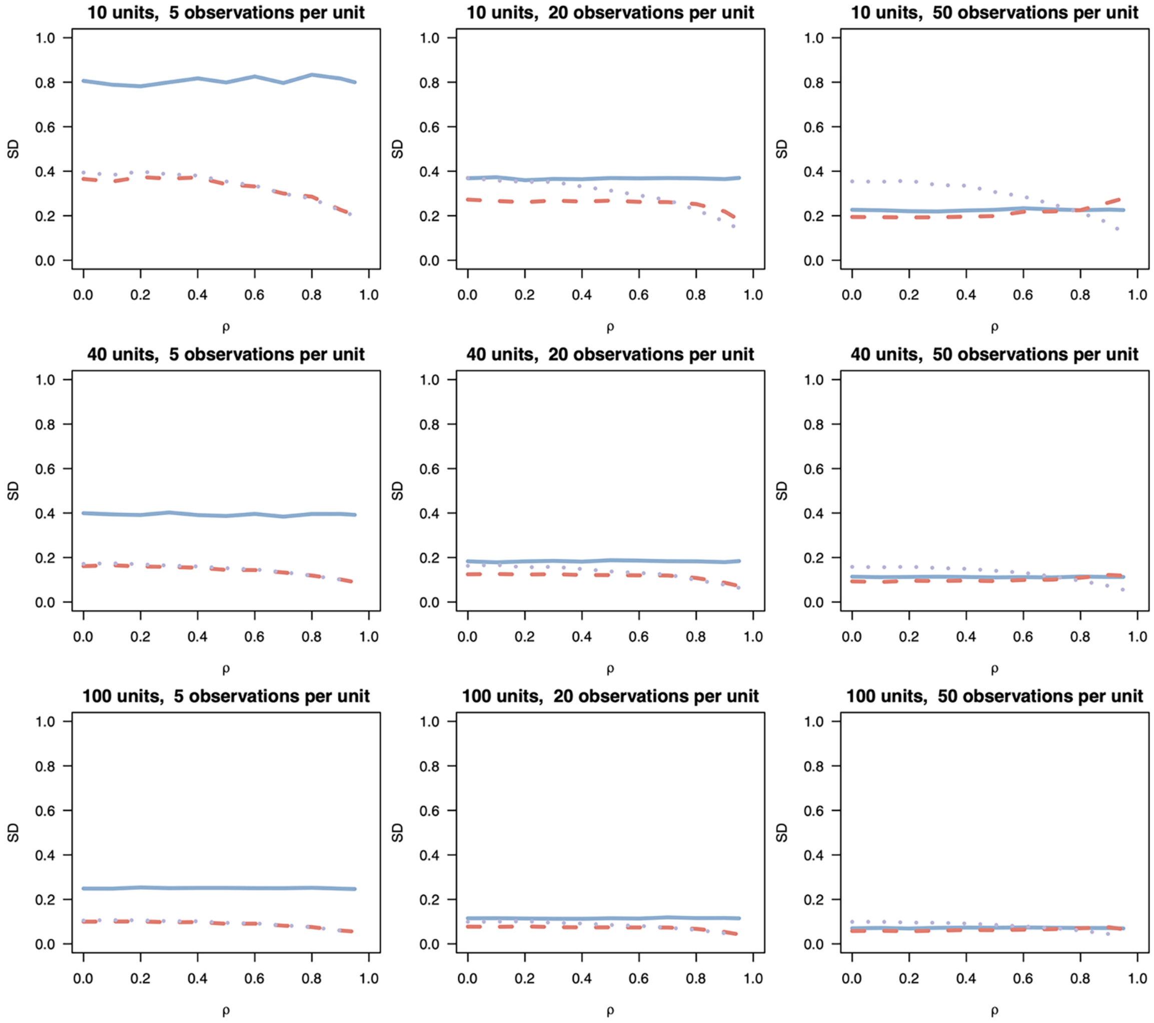

Figure 8. SD of $\hat{\beta}$, Clark and Linzer's sluggish case. Solid line = OLS-FE, dashed line = FGLS-RE, dotted line = OLS-pooled model; the horizontal axis is the value of correlation between $\bar{x}_j$ and α_j (ρ).

effect of the independent variable of four time series models: EQ4 (an ADL(2,1) specification, given in eq. [4] of Wilkins [2018]), LGDV (a lagged dependent variable model), LGDV2 (a lagged dependent variable model with two lags of the DV), and a static model.[22] From his results using only percentage bias and RMSE, Wilkins concludes that LGDV is the preferred model at low levels of autocorrelation and EQ4 is the preferred model at higher levels of autocorrelation. From our extension of his analysis, we come to fairly different conclusions.

From the evaluation stage, we find that EQ4 has the highest RMSE at low levels of autocorrelation (less than 0.3). When there is no autocorrelation, all models have the expected value of coverage probability (0.95). However, as the amount of autocorrelation increases, the coverage probability of the LGDV and LGDV2 models decreases, while that of EQ4 remains around 0.95. All models have high power. From our diagnoses, we find that the higher RMSE values of EQ4 at low levels of autocorrelation (less than 0.3) are because of its higher SD and that the higher RMSE values of the LGDV and LGDV2 models at higher levels of autocorrelation (above 0.3) are mainly because of its bias, which is not offset by its lower SD. EQ4's expected coverage probabilities are a result of its unbiasedness and ability to recover accurate standard errors, despite having a relatively high SD. The lower coverage probabilities for the LGDV and LGDV2 models are due to a combination of bias and overconfidence. The high power for the LGDV and LGDV2 models is a result of their overconfidence, despite being biased toward the false null hypothesis ($\beta = 0$). Overall, we come to a more nuanced conclusion than that of Wilkins: in terms of point estimates, the LGDV and LGDV2 models are preferred at low levels of autocorrelation, and EQ4 is almost always preferred in terms of inference.

[22] We omit the results from the static model because of its extremely poor performance.

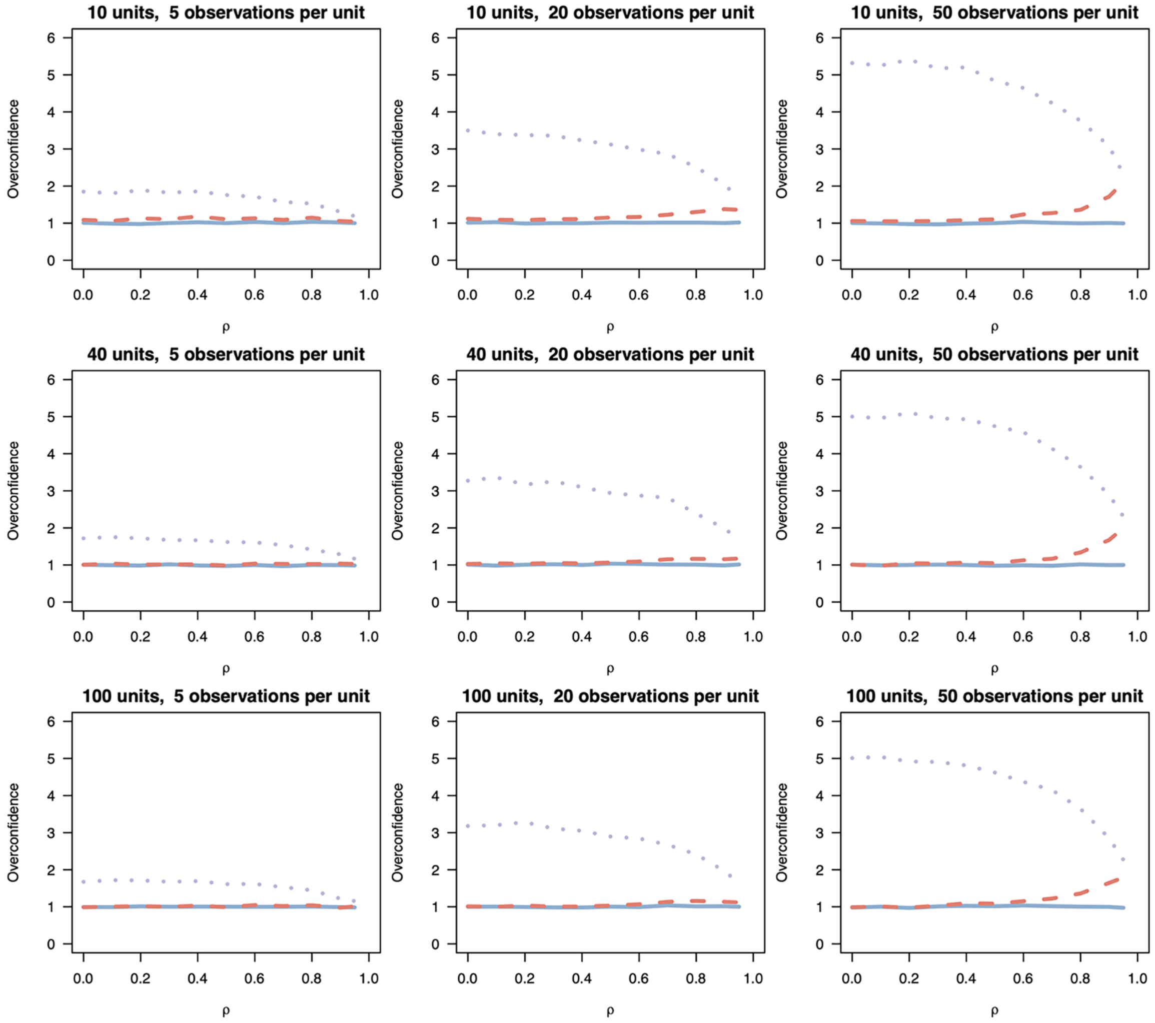

Figure 9. Overconfidence of $\hat{\beta}$, Clark and Linzer's sluggish case. Solid line = OLS-FE, dashed line = FGLS-RE, dotted line = OLS-pooled model; the horizontal axis is the value of correlation between $\bar{x}_j$ and α_j (ρ).

Hanmer and Kalkan (2013) compare the performances of the average marginal effects (AME) and marginal effects at means (MEM) approaches for probit models in the presence of omitted variables.[23] They compare these marginal effects when one covariate is excluded to those when the model is correctly specified and find that the AME approach is preferred because of its unbiasedness.[24] In evaluating these two approaches, our results demonstrate that the AME approach has lower RMSE values, close-to-expected coverage probability, and a power of 1. The MEM approach, however, has a low coverage probability and a power of 1. In diagnosing these performances, we find that the lower RMSE values of the AME approach are due to a combination of its unbiasedness and lower SD. The AME approach recovers close-to-expected levels of coverage probability because, while both approaches perform similarly in recovering accurate standard errors (overconfidence close to 1), the AME approach is unbiased. The low coverage probability of the MEM approach, despite its higher SD, is because of its bias, which also contributes to its high power. Across the board, the AME approach is preferred.

In table 2, we summarize the results of all three replications and extensions. In the case of both Clark and Linzer (2015) and Wilkins (2018), we find that our conclusions differ substantially from those of the original studies. In the case of Hanmer and Kalkan (2013), we arrive at the same conclusion as the original study but demonstrate that their conclusions are robust to our recommended considerations of estimator quality.

[23] Both AME and MEM approaches have been used to obtain what is a typical effect of a shift in an independent variable on predicted probabilities from probit and logit models. Although they can be thought of as different quantities of inference for users of such models, the goal of the authors is to compare the performance of these two rival estimators of typical effects and their sensitivity to omitted variable bias.

[24] In this paragraph, we only discuss the replication of model 1, panel A, table 1 in Hanmer and Kalkan (2013). The entire replication is provided in the appendix.

Table 2. Summary of Our Replications and Extensions

	Original	Replication and Extension
Clark and Linzer (2015):		
Performance statistic	R	R C P B S O
Conclusion	When the correlation between unit effects and the predictor, within-unit variation, and the number of within-unit observations are all low, RMSE demonstrates the RE estimator is better than the FE estimator	Different sample sizes and levels of correlation influence whether FE or RE performs better in terms of point estimates, but the FE estimator always performs better for inference unless the correlation between unit effects and the predictor is 0
Wilkins (2018):		
Performance statistic	B R	R C P B S O
Conclusion	When both the dependent and independent variables are highly autoregressive, the EQ4 model has lower bias. At higher levels of serial autocorrelation, the EQ4 model performs better in terms of RMSE	The LGDV and LGDV2 models perform better for point estimates at low levels of autocorrelation, and EQ4 at higher levels. With regard to inference, EQ4 almost always performs best
Hanmer and Kalkan (2013):		
Performance statistic	B	R C P B S O
Conclusion	The AME approach is preferable to the MEM approach because it produces less biased marginal effects estimates when relevant variables are omitted	For the covered circumstances, the AME approach is always preferred

Note. Letters indicate which performance statistics were reported in the original study and in our replication. R = RMSE, C = coverage probability, P = power, B = bias, S = SD, O = overconfidence.

CONCLUSION

Articles that report the results of Monte Carlo experiments play an important role in political science. They disperse knowledge about new statistical techniques and estimator properties and serve as references for scholars interested in using these estimators to test their theoretical expectations. Given that a substantial amount of research in political science is shaped by such recommendations, these decisions should be based on the most important dimensions of estimator performance. Reasonable people can, of course, disagree about the relative importance of different performance statistics.

As we mention in the introduction, our article is designed to help two audiences. For those who produce Monte Carlo simulations, we offer guidance about which performance statistics to report. We identify patterns in what gets reported (as well as what does not) and show how combining statistics can improve analysis. For those who read Monte Carlo work, we provide a useful overview of the six most common performance statistics in order to help readers think critically and systematically about the results from these simulations.

With this in mind, we present a new way to think about the advantages of the different performance statistics, both independently and in combination. For the purposes of evaluating point estimates, we encourage comparing performances by examining the RMSE. For the purposes of evaluating inference, we encourage comparing performances in terms of coverage probability and power. We believe these three performance statistics—RMSE, coverage probability, and power—are of most use to researchers interested in knowing which method to use because they provide information about the average error of a method as well as the ability to make accurate hypothesis-testing inferences. We also recommend that researchers who want to diagnose the source of performance (good or bad) use combinations of bias, SD, and overconfidence.

ACKNOWLEDGMENTS

We thank Kentaro Fukumoto, Adam Glynn, Paul Kellstedt, Vera Troeger, and Laron Williams for their helpful comments on earlier versions of this manuscript. We also appreciate the comments made by those attending presentations of earlier versions of this project at the annual meetings of the Midwest Political Science Association (2017, 2018), meetings of the Dynamic Pie group (2018, 2019), the Texas A&M European Union Center's conference Modeling Politics and Policy in Time and Space (2017), and the Annual Texas Methods (TexMeth) Meeting (2017).

REFERENCES

Barbu, Adrian, and Song-Chun Zhu. 2020. *Monte Carlo Methods*. Dordrecht: Springer.

Beck, Nathaniel, and Jonathan N. Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data." *American Political Science Review* 89 (3): 634–47.

Carsey, Thomas M., and Jeffrey J. Harden. 2014. *Monte Carlo Simulation and Resampling Methods for Social Science*. 1st ed. Thousand Oaks, CA: Sage.

Clark, Tom S., and Drew A. Linzer. 2015. "Should I Use Fixed or Random Effects?" *Political Science Research and Methods* 3 (2): 399–408.

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

Franzese, Robert J., and Jude C. Hays. 2007. "Spatial Econometric Models of Cross-Sectional Interdependence in Political Science Panel and Time-Series-Cross-Section Data." *Political Analysis* 15 (2): 140–64.

Gill, Jeff. 2014. *Bayesian Methods: A Social and Behavioral Sciences Approach*, vol. 20. Washington, DC: CRC Press.

Hanmer, Michael J., and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57 (1): 263–77.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Dordrecht: Springer.

Helgason, Agnar Freyr. 2016. "Fractional Integration Methods and Short Time Series: Evidence from a Simulation Study." *Political Analysis* 24 (1): 59–68.

Honaker, James, Jonathan N. Katz, and Gary King. 2002. "A Fast, Easy, and Efficient Estimator for Multiparty Electoral Data." *Political Analysis* 10 (1): 84–100.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*, vol. 846. Hoboken, NJ: Wiley.

Keele, Luke, and Nathan J. Kelly. 2006. "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables." *Political Analysis* 14 (2): 186–205.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (2): 137–63.

Philips, Andrew Q. 2018. "Have Your Cake and Eat It Too? Cointegration and Dynamic Inference from Autoregressive Distributed Lag Models." *American Journal of Political Science* 62 (1): 230–44.

Philips, Andrew Q. 2022. "How to Avoid Incorrect Inferences (While Gaining Correct Ones) in Dynamic Models." *Political Science Research and Methods* 10 (4): 879–89.

Pickup, Mark, and Vincent Hopkins. 2022. "Transformed-Likelihood Estimators for Dynamic Panel Models with a Very Small *T*." *Political Science Research and Methods* 10 (2): 333–520.

Plümper, Thomas, and Vera E. Troeger. 2007. "Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects." *Political Analysis* 15 (2): 124–39.

Plümper, Thomas, and Vera E. Troeger. 2019. "Not So Harmless after All: The Fixed-Effects Model." *Political Analysis* 27 (1): 21–45.

Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58 (4): 1083–91.

Robert, Christian P., and George Casella. 2010. *Monte Carlo Statistical Methods*. 2nd ed. New York: Springer.

Thomopoulos, Nick T. 2012. *Essentials of Monte Carlo Simulation: Statistical Methods for Building Simulation Models*. Dordrecht: Springer.

Webb, Clayton, Suzanna Linn, and Matthew J. Lebo. 2020. "Beyond the Unit Root Question: Uncertainty and Inference." *American Journal of Political Science* 64 (2): 275–92.

Whitten, Guy D., Laron K. Williams, and Cameron Wimpy. 2021. "Interpretation: The Final Spatial Frontier." *Political Science Research and Methods* 9 (1): 140–56.

Wilkins, Arjun S. 2018. "To Lag or Not to Lag? Re-evaluating the Use of Lagged Dependent Variables in Regression Analysis." *Political Science Research and Methods* 6 (2): 393–411.