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

Vincent Hopkins†
Ali Kagalwala‡
Andrew Q. Philips§
Mark Pickup¶
Guy D. Whitten||

Abstract

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 and/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.
Keywords: Monte Carlo simulations; RMSE; Bias; Coverage probability; Power; Standard Deviation, Overconfidence.

*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.

†Vincent Hopkins is an Assistant Professor in the Department of Political Science, University of British Columbia, Vancouver, BC V6T 1Z1. vince.hopkins@ubc.ca

‡Ali Kagalwala is a PhD Candidate in Political Science in The Bush School of Government and Public Service, Texas A&M University, College Station, TX 77843. alikagalwala@tamu.edu

§Andrew Q. Philips is an Associate Professor in the Department of Political Science, University of Colorado Boulder, Boulder, CO 80309. Andrew.Philips@colorado.edu

¶Mark Pickup is a Professor in the Department of Political Science at Simon Fraser University, Burnaby, BC, Canada V5A 1S6. mark.pickup@sfu.ca

||Guy D. Whitten 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. g-whitten@tamu.edu

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; King and Zeng, 2001; Keele and Kelly, 2006; 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, overconfidence, coverage, and power often should also be reported. Doing so will allow researchers to make more informed decisions about which method(s) 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 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 standard deviation. We provide a way to think through what can be learned from various combinations—e.g., if an estimator shows no evidence of bias, we explain what might then be gleaned from the standard deviation. Our paper also helps the second group, readers of Monte Carlo work, to better

understand the tradeoffs 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 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, Tibshi-

[1]For general overviews of Monte Carlo methods, c.f., Barbu and Zhu (2020) or Tho-

rani 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 and/or interpretation of statistical models (e.g., Gill, 2014; Jackman, 2009). In this paper, 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, standard deviation, overconfidence, RMSE, coverage, and power). For each performance statistic, we provide a definition, the relevant formulae (if needed), and a short summary of the statistic's importance.

Bias, standard deviation, and overconfidence

mopoulos (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, 2019).

[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."

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

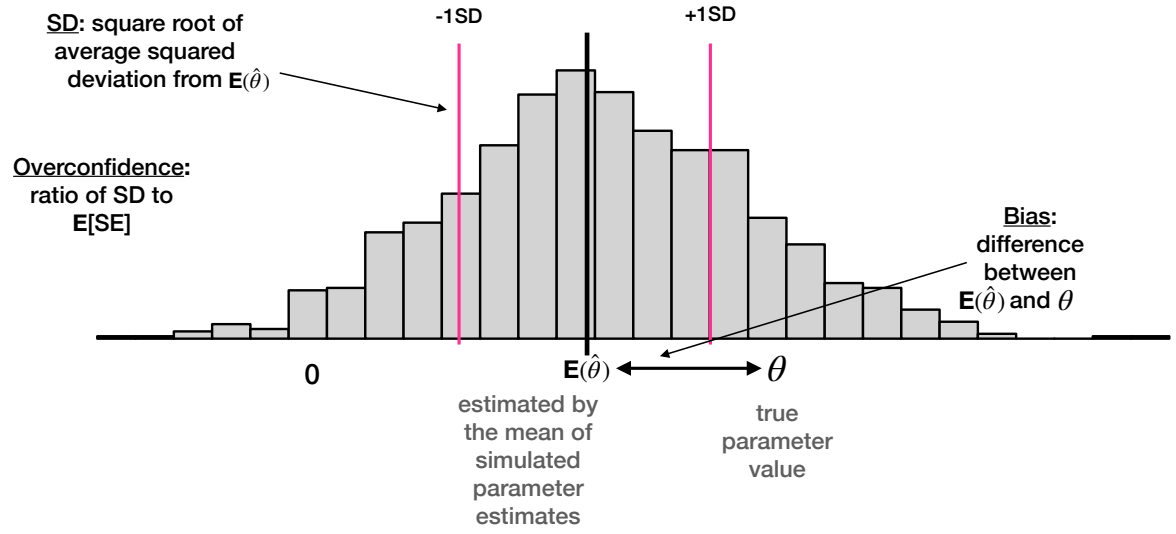

Figure 1: Illustration of bias, SD, and overconfidence

Bias

Definition and formulae: 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 quantity from repeated sampling and the value of the quantity in the DGP. When $\mathrm{E}(\hat{\theta}) \neq \theta$, as in the figure, the estimator is biased.

$$\textbf{Definition: } \mathrm{Bias}[\hat{\theta}] = \mathrm{E}[\hat{\theta} - \theta] = \mathrm{E}[\hat{\theta}] - \theta \tag{1}$$

$$\textbf{Calculation: } \widehat{\mathrm{Bias}}[\hat{\theta}] = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta) \tag{2}$$

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 "aver-

age bias" is by far the most commonly calculated quantity, others are possible, including median bias (c.f., Pickup and Hopkins, 2020), which is useful when the distribution of the quantity of interest is not normally distributed (e.g., when calculating non-linear 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 (c.f., 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 standard deviation (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 $\mathrm{E}(\hat{\theta})$.

$$\textbf{Definition: } \mathrm{SD}[\hat{\theta}] = \sqrt{\mathrm{E}[(\hat{\theta} - \mathrm{E}[\hat{\theta}])^2]} \tag{3}$$

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

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 $\mathrm{E}[\hat{\theta}]$, which itself may or may not

[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 which, all else equal, has the lowest bias. For another discussion of the importance of bias, see Carsey and Harden (2014).

be biased (e.g., $\mathrm{E}[\hat{\theta}]$ 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 standard deviation of the estimates divided by the expected value of the estimated standard errors for a quantity of interest.[5]

$$\textbf{Definition: } \text{Overconfidence}(\hat{\theta}) = \frac{\mathrm{SD}(\hat{\theta})}{\mathrm{E}\big[\text{s.e.}(\hat{\theta})\big]} \tag{5}$$

$$\textbf{Calculation: } \widehat{\text{Overconfidence}}(\hat{\theta}) = \frac{\widehat{\mathrm{SD}}[\hat{\theta}]}{\frac{1}{n}\sum_{i=1}^{n}\text{s.e.}(\hat{\theta}_i)} \tag{6}$$

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

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

[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.

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

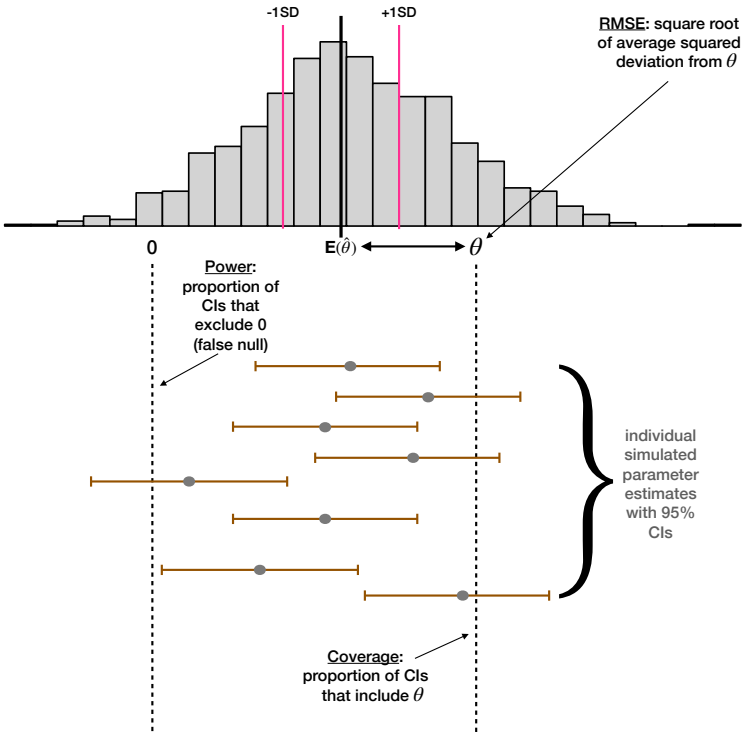

Figure 2: Illustration of RMSE, coverage, and power

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 gray bars. The dotted line on the left side of this figure shows the value of the false null hypothesis, specified as zero, and the dotted line on the right side of this figure shows

the DGP value of θ. The horizontal confidence intervals show a series of results from hypothetical simulations for an estimator, $\hat{\theta}$.[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: The root mean squared error 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.

$$\textbf{Definition: } \text{RMSE}[\hat{\theta}] = \sqrt{\text{E}[(\hat{\theta} - \theta)^2]} = \sqrt{\text{Bias}(\hat{\theta})^2 + \text{SD}^2(\hat{\theta})} \tag{7}$$

$$\textbf{Calculation: } \widehat{\text{RMSE}}[\hat{\theta}] = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(\hat{\theta}_i - \theta)^2\right]} \tag{8}$$

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 tradeoff. E.g., 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

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

in accuracy from one estimator are larger than those from other estimators.[9]

Coverage Probability

Definition: As we illustrate in Figure 2, coverage probability is the proportion of times the confidence intervals of the estimator encompasses the true DGP value. It is calculated as the proportion of simulated 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 dotted 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 they are 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(Type 1 error) $= 1-$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 hypoth-

[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 Sections 3 and 4 below for a discussion of why calculating all three performance statistics is important.

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

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

esis 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. 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 simulated 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 dotted 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(Type 2 error) $= 1-$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, and coverage probability and power—*evaluates* an estimator's performance on point estimates and inference. The second group

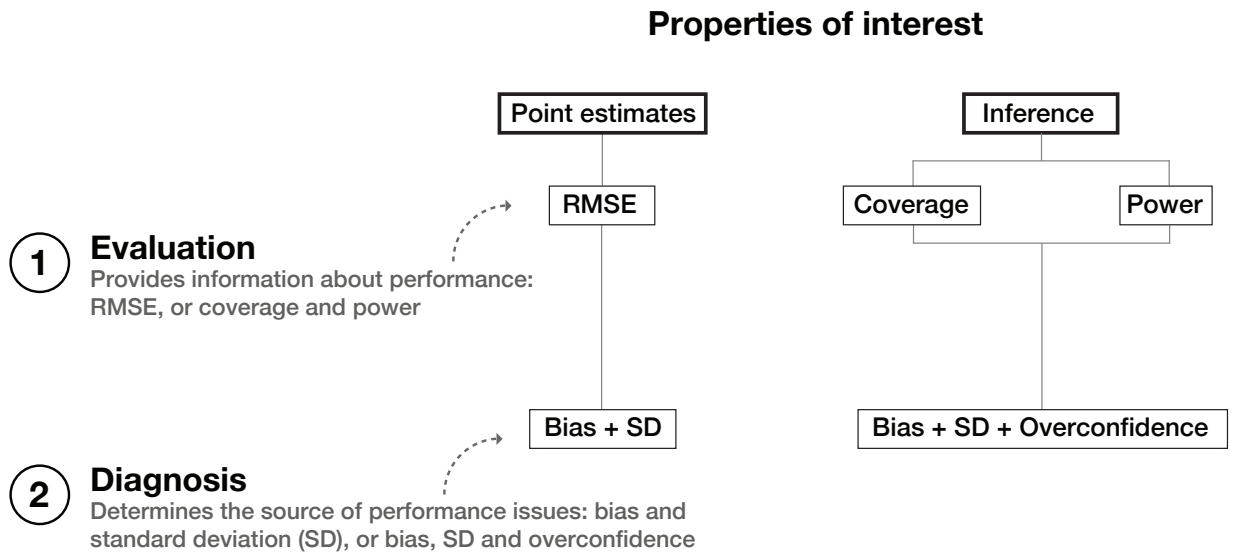

Figure 3: Information provided by performance statistics

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

[12]This does not necessarily mean starting with RMSE. E.g., if a study compares the performance of different robust standard errors (SE) 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 SEs. We also note that if there is poor coverage and/or power, then overconfidence (and SD) can shed light on why this is the case. On the other hand, if simulations show no problems with power and coverage (or they are good enough that we are comfortable with the performance of the robust SE), then we can be confident there are no problems with overconfidence.

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 due to the systematic over- or under-estimation 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 if 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

[13]By "small," we generally mean close enough to 0 that we expect estimates to be within the precision of our original measures.

are substantially different; again, whether bias and SD differ across estimators is hard to know without directly calculating these two performance measures. 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, e.g. 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 if 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. E.g., 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

[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.

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 and/or overconfidence should increase power but the latter does so by incorrectly estimating the precision of the estimate. Holding all else constant, attenuation bias ($0 < |\mathrm{E}[\theta]| < |\theta|$) 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 ($0 < |\theta| < |\mathrm{E}[\theta]|$) 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 and/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 that perform poorly and differently with respect to point estimates (RMSE) and/or inference (coverage probability and/or power). If the estimators perform well and/or similarly on one or more performance statistics, those results need only a brief mention.
2. Diagnose the estimators that perform poorly and/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 and/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. E.g., consider Philips (2021), who generates two independent unit roots in one of his Monte Carlo experiments and compares the performance of three time series model in terms of Type 1 error rates of the long-run effects—LDV, ECM, and ADL(1,1).[15] He finds all three models perform similarly, and poorly, in terms of the coverage probability of the long-run effect. Based on our recommendations, he should summarize the results for coverage probability briefly in text, e.g., "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 and/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 appendices 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 paper, when examining all six performance statistics, we can learn novel and important things about estimators that may lead to different conclusions about the preferred estimator than those of the original author(s). Before turning to these replications, we present a review of 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*,

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

and *Political Science Research and Methods* from 2006 to 2016 that contained the keywords "Monte Carlo" and/or "simulation."[16]

Performance Statistic							
Bias	RMSE or MSE	Coverage/ Type 1	SD	Over-confidence	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	

Table 1: The most common patterns of reporting performance statistics in major Political Science journals

Notes: The letters in each row indicate that that 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 SM for the full table of reporting patterns.

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 between the two horizontal lines 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

[16]Since publication of *Political Science Research and Methods* began in 2013, we coded 2013-2016 for that journal. We coded all Monte Carlo simulations that were presented as a part of published papers and in appendices that appeared as a part of the volume in which they were published; see the SM for details.

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

present in 85.9 percent 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 (c.f., Honaker, Katz and King, 2002; Esarey, 2016; Helgason, 2016) visually show more than one quantity of interest—for instance bias 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 our findings and provide full details in the SM.

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" (p. 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:

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

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

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

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. ϵ_i is an i.i.d. error term 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 zero, variance of one, and covariance between $\bar{x}_j$ and α_j, equal to ρ ($\rho = 0, 0.1, 0.2 \cdots, 0.9, 0.95$). 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), when the number of within-unit observations (n) is small, and the amount of correlation between the unit intercepts and 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 2000 datasets 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 (p. 406). As is clear from the figure, the OLS-FE estimator (the blue 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 (the red dashed line) and the OLS-pooled (purple 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

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

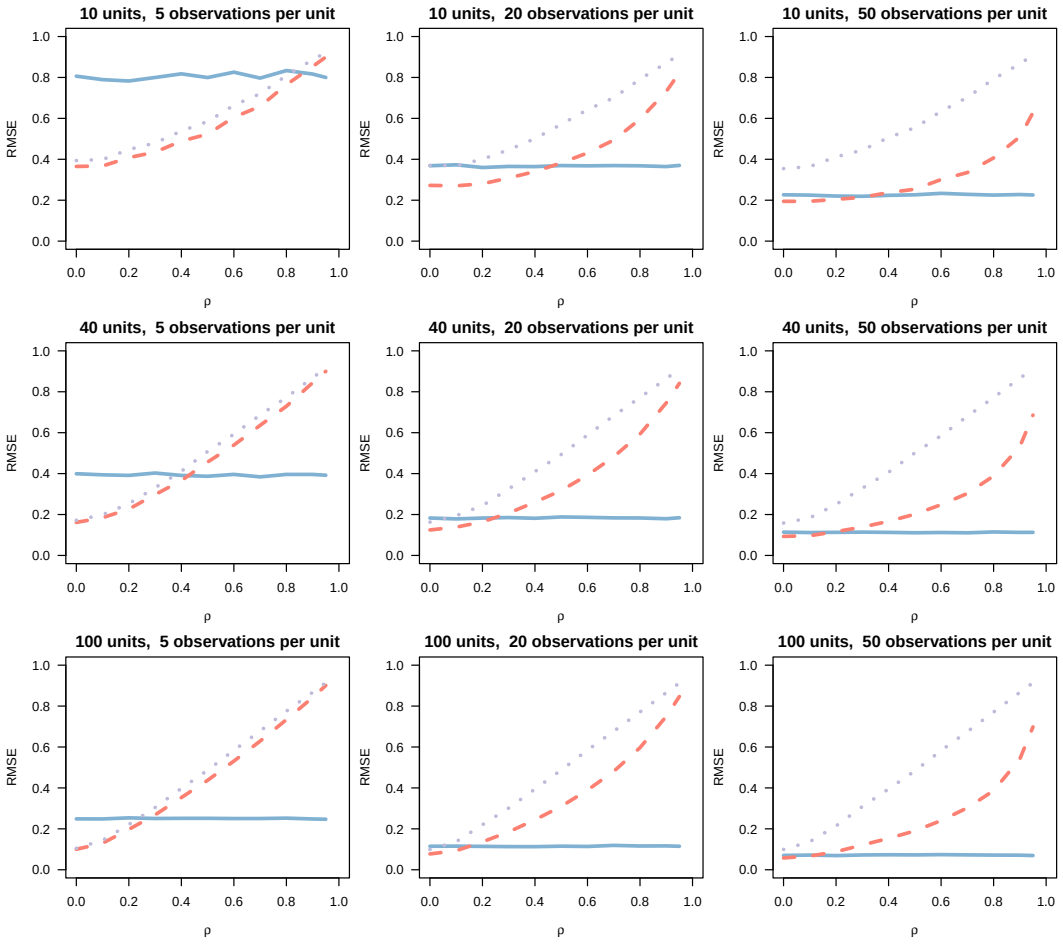

Figure 4: RMSE of $\hat{\beta}$, Clark and Linzer's sluggish case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

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.

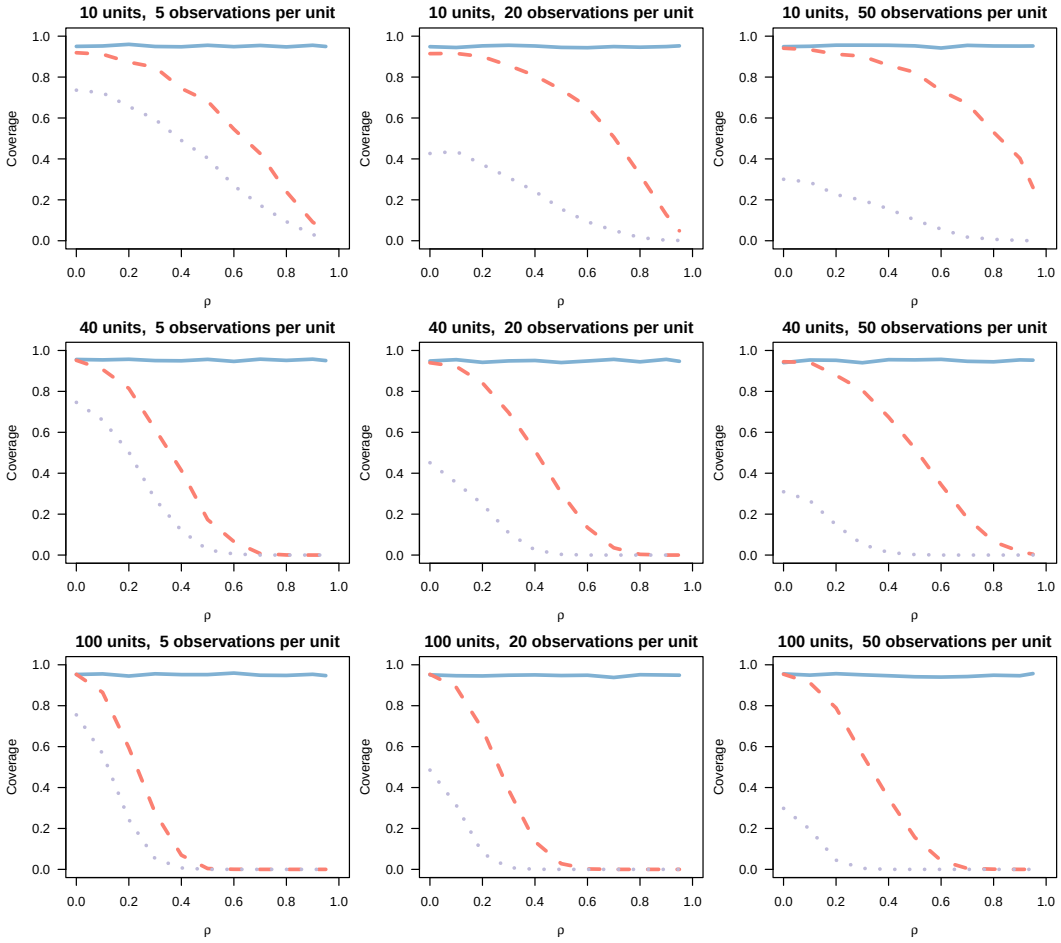

Figure 5: Coverage probability of $\hat{\beta}$, Clark and Linzer's sluggish case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

In Figure 5, we show the coverage probability statistics of the estimators (i.e., how often the 95 percent 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 non-zero, 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.

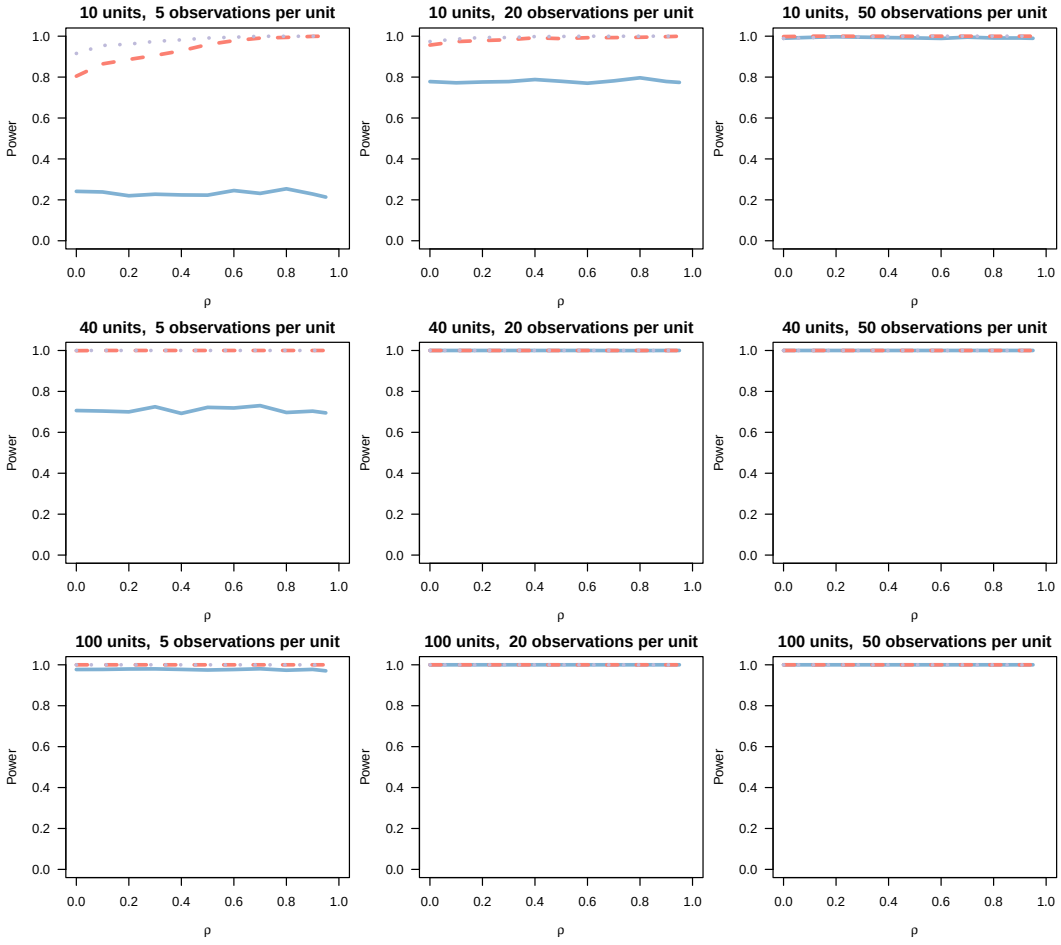

Figure 6: Power of $\hat{\beta}$, Clark and Linzer's sluggish case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

In Figure 6 we consider the power of the estimators; i.e., 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.

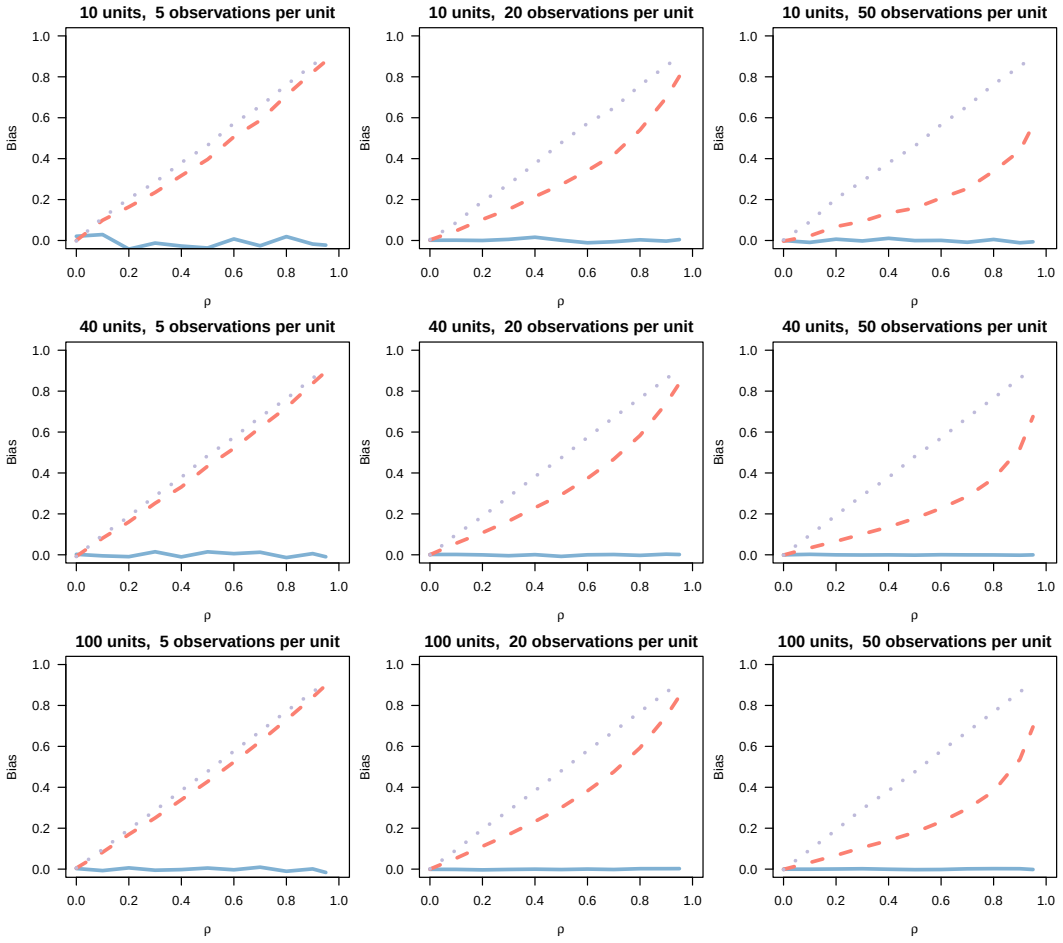

Figure 7: Bias of $\hat{\beta}$, Clark and Linzer's sluggish case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

As noted in Figure 3, RMSE, coverage probability, and power merely evaluate esti-

mators' performances with respect to point estimates and inferences, but 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 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 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 non-zero value of ρ and this bias increases as the value of ρ increases. Only when $\rho = 0$ and there are 10 unit and 5 within-unit observations do the pooled and random effects estimators perform better than fixed effects, and only by a small amount. For any non-zero 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 standard deviations 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 Figure 8). These relative gains in efficiency decrease as both J and n increase, and the standard deviations 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 standard deviation 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.

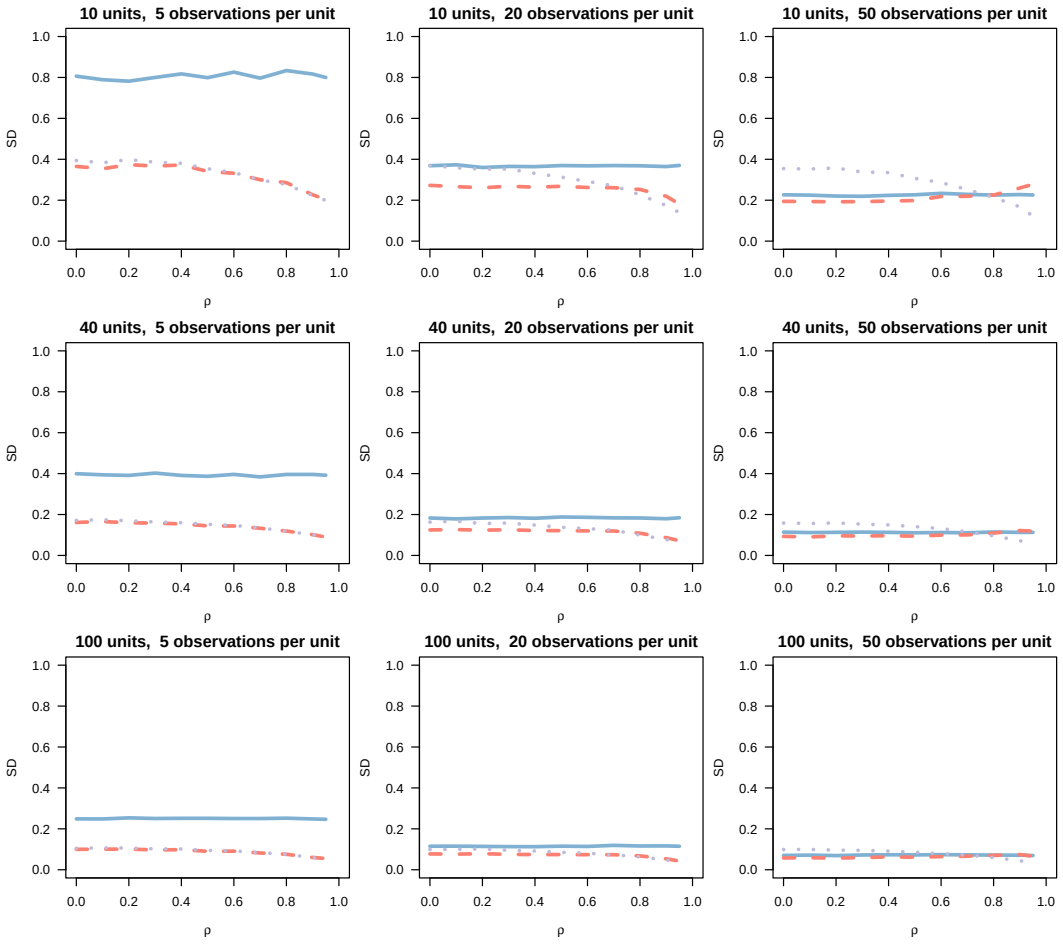

Figure 8: Standard deviation of $\hat{\beta}$, Clark and Linzer's sluggish case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

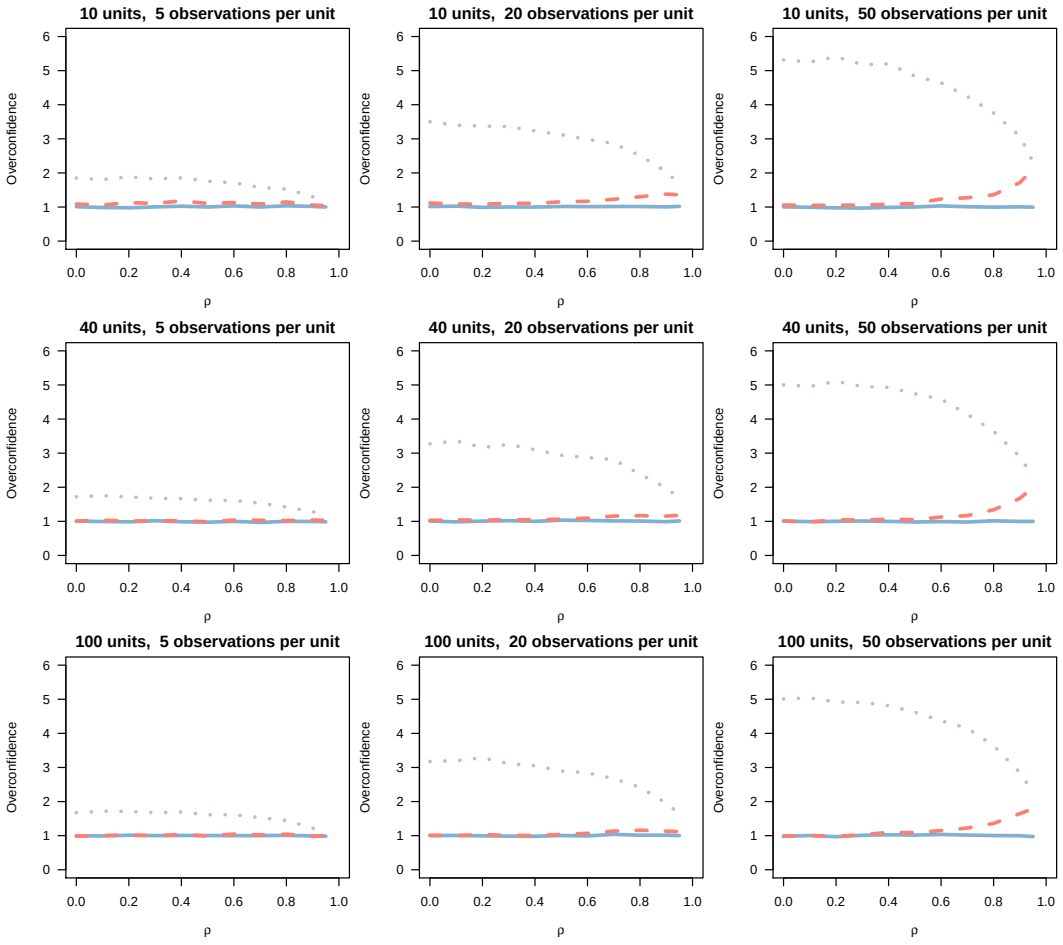

Figure 9: Overconfidence of $\hat{\beta}$, Clark and Linzer's sluggish case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

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 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 in Section , this is an assessment of whether the overconfidence measure differs from one. 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 standard deviation 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 standard deviation 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 standard deviation 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

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

intervals due to its SD, making it more likely 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 which 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, standard deviation, 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.

Clark and Linzer (2015, p.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

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

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 $\rho = 0$ (though, 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 SM 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 percent bias and average error (RMSE) in the short-run effect of the independent variable of four time series models: EQ4 (an ADL(2,1) specification), 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 percent 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.

[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.

[22]We omit the results from the static model due to its extremely poor performance.

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 model 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 are a result of their overconfidence, despite being biased towards 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.

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

[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

demonstrate that the AME approach has lower RMSE values, close-to-expected coverage probability, and a power of 1. The MEM approach, on the other hand, has a low coverage probability and a power of 1. In diagnosing these performances, we find that the lower RSME 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.

Authors	**Original**	**Replication and extension**
Clark and Linzer (2015)	*Performance statistics*: R *Conclusions*: 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.	*Performance statistics*: R C P B S O *Conclusions*: 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 statistics*: B R *Conclusions*: 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.	*Performance statistics*: R C P B S O *Conclusions*: The LGDV and LGDV2 models perform better for point estimates at low levels of autocorrelation, and EQ4 at higher levels. With regards to inference, EQ4 almost always performs best.
Hanmer and Kalkan (2013)	*Performance statistics*: B *Conclusions*: The AME approach is preferable to the MEM approach because it produces less biased marginal effects estimates when relevant variables are omitted.	*Performance statistics*: R C P B S O *Conclusions*: For the covered circumstances, the AME approach is always preferred.

Table 2: Summary of our replications and extensions

Note: The 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.

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

Hanmer and Kalkan (2013). The entire replication is provided in the SM.

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.

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 paper 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 interest 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 performances (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 Meeting 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 on "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*. Springer.

Beck, Nathaniel and Jonathan N. Katz. 1995. "What to do (and Not to Do) with Time-Series Cross-Section Data." *The American Political Science Review* 89(3):634–647.

Carsey, Thomas M. and Jeffrey J. Harden. 2014. *Monte Carlo Simulation and Resampling Methods for Social Science*. First ed. Thousand Oaks: SAGE Publications.

Clark, Tom S and Drew A Linzer. 2015. "Should I use fixed or random effects?" *Political Science Research and Methods* 3(02):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):140–164.

Gill, Jeff. 2014. *Bayesian methods: A social and behavioral sciences approach*. Vol. 20 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):263–277.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

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 John Wiley & Sons.

Keele, Luke and Nathan J Kelly. 2006. "Dynamic models for dynamic theories: The ins and outs of lagged dependent variables." *Political analysis* pp. 186–205.

King, Gary and Langche Zeng. 2001. "Logistic regression in rare events data." *Political analysis* 9(2):137–163.

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):230–244.

Philips, Andrew Q. 2021. "How to avoid incorrect inferences (while gaining correctones) in dynamic models." *Political Science Research and Methods* pp. 1–11.

Pickup, Mark and Vincent Hopkins. 2020. "Transformed-likelihood estimators for dynamic panel models with a very small T." *Political Science Research and Methods* pp. 1–20.

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.

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):1083–1091.

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

Thomopoulos, Nick T. 2012. *Essentials of Monte Carlo simulation: Statistical methods for building simulation models*. Springer Science & Business Media.

Webb, Clayton, Suzanna Linn and Matthew J Lebo. 2020. "Beyond the Unit Root Question: Uncertainty and Inference." *American Journal of Political Science* 64(2):275–292.

Whitten, Guy D, Laron K Williams and Cameron Wimpy. 2019. "Interpretation: the final spatial frontier." *Political Science Research and Methods* pp. 1–17.

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.

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

Vincent Hopkins*
Ali Kagalwala†
Andrew Q. Philips‡
Mark Pickup§
Guy D. Whitten¶

*Vincent Hopkins is an Assistant Professor in the Department of Political Science, University of British Columbia, Vancouver, BC V6T 1Z1. vince.hopkins@ubc.ca

†Ali Kagalwala is a PhD Candidate in Political Science in The Bush School of Government and Public Service, Texas A&M University, College Station, TX 77843. alikagalwala@tamu.edu

‡Andrew Q. Philips is an Associate Professor in the Department of Political Science, University of Colorado Boulder, Boulder, CO 80309. Andrew.Philips@colorado.edu

§Mark Pickup is a Professor in the Department of Political Science at Simon Fraser University, Burnaby, BC, Canada V5A 1S6. mark.pickup@sfu.ca

¶Guy D. Whitten 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. g-whitten@tamu.edu

Appendix A. Measures of Overconfidence

Our equation for calculating overconfidence is the same equation as used by Franzese and Hays (2007, p. 154), except that we have flipped the numerator and denominator (we did this because larger values representing greater overconfidence seems more intuitive). In this way, our equation is similar to Beck and Katz (1995), although our equation is different from theirs. If the two equations are put in the same notation, here's how they compare:

$$\textbf{Beck and Katz (1995):}\ \widehat{\text{Overconfidence}}(\hat{\theta}) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\hat{\theta}}_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\text{s.e.}(\hat{\theta}_i^2)}} \tag{1}$$

$$\textbf{Us:}\ \widehat{\text{Overconfidence}}(\hat{\theta}) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\hat{\theta}}_i)^2}}{\frac{1}{n}\sum_{i=1}^{n}\text{s.e.}(\hat{\theta}_i)} \tag{2}$$

While the numerators are identical, the denominators are not. Consequently, we are comparing the standard deviation of θ (the numerator) to the average calculated standard error for θ (the denominator), while Beck and Katz (1995) are comparing the standard deviation of θ to the square root of the average squared standard error (i.e., estimated variance). We believe that the correct comparison is to the average standard error (as is done in our measure), while taking the square root of the average estimated variance, as Beck and Katz do, is not the same thing. Since Beck and Katz square all of the standard errors, take the average, and then the square root, the resulting value is larger than the average of the original set of standard errors. As a consequence the Beck and Katz measure of overconfidence is deflated. We think that it is also likely that as the standard deviation of the SEs increases, the absolute difference between the average SE on one hand and the square root of the average squared SE on the other increases.

To better see this, we conducted a Monte Carlo experiment with the following DGP:

$$y_i = \beta x_i + \epsilon_i,\ \text{where}\ \epsilon_i \sim N(0, \sigma^2) \tag{3}$$

Where we set $\beta = 2$, varied σ^2, the number of observations in the DGP (N) and the number of Monte Carlo simulations, n. We obtained $\hat{\beta}$ as well as its standard error across the n simulations, and calculated each overconfidence measure. Table 1 shows our results, as well as the calculated average standard error, the standard deviation of the n standard errors, the average of $\hat{\beta}$ and the standard deviation of the $\hat{\beta}$'s. Overconfidence measures should show any discrepancies between the average standard error and the standard deviation of the estimates themselves; if the average standard error is larger (smaller) than the standard deviation of the $\hat{\beta}$'s, this means that the standard errors are too wide (small), and thus underconfident (overconfident). While both our measure and Beck and Katz's are often close, they are not always the same. Moreover, there are times in which the Beck and Katz measure gets over/under-confidence wrong. Consider the case where $N = 100$, $n = 100,000$ and $\sigma^2 = 1$; the average SE from these simulations is 0.1009591, which is *smaller* than the SD of the estimates, 0.10133717. While our measure of 100.37 correctly concludes that the standard errors are overconfident (i.e., $\bar{\text{S.E.}}$ <SD of $\hat{\beta}$), the

Beck and Katz measure of 99.86 would lead us to incorrectly conclude that the standard errors were too large.

N	n	σ^2	S.$\bar{\text{E}}$.	SD of SE's	$\bar{\hat{\beta}}$	SD of $\hat{\beta}$'s	O/U Conf?	Beck-Katz	Ours
100	100	1	0.1012794	0.01032475	2.006	0.09812674	U	95.90936	96.40148
100	1000	1	0.1011460	0.01032285	2.004	0.10091640	U	99.20827	99.72310
100	100000	1	0.1009591	0.01023502	1.999	0.10133717	O	99.86214	100.37399
100	100	10	1.0083532	0.09781574	2.043	1.17172661	O	115.08472	115.61954
100	1000	10	0.9991330	0.10115902	1.985	0.98883985	U	98.41765	98.92030
100	100000	10	1.0104994	0.10214676	1.998	1.01539605	O	99.97460	100.48408
25	100	10	2.0783445	0.39428782	1.877	2.35406058	O	110.74272	112.69839
25	1000	10	2.0755450	0.43711981	1.850	2.19089345	O	103.24217	105.50471
25	100000	10	2.0831946	0.44559429	2.007	2.13951190	O	100.43110	102.70290

Table 1: Overconfidence Monte Carlo results

Note: U = underconfident, S.$\bar{\text{E}}$. >SD of $\hat{\beta}$. O = overconfident, S.$\bar{\text{E}}$. <SD of $\hat{\beta}$

Appendix B. Coverage, Power, and Hypothesis Test Specification

As we discuss in the paper, coverage and power performance measures are each partially determined by a series of test specification choices made by the researcher. These include the specification of a null hypothesis, alternative hypothesis, and significance level. Here we provide some further details on the relationship between coverage, power, and hypothesis test specification.

A consistent estimator is defined as one for which, as the sample size tends to infinity, the estimate converges on the true parameter value and its variance approaches 0. The coverage probability of a consistent estimator depends on the sample size and the significance level (e.g. 1%, 5%, 10%). In finite samples, the expected coverage probabilities for the 10%, 5%, and 1% significance levels are 0.90, 0.95, and 0.99 respectively.

The power of a consistent estimator depends on the null hypothesis, sample size, and significance level (Greene, 2017). For instance, if the null hypothesis is $\theta = 2$ and the true parameter value is 2.1, it is likely the estimator will be unable to reject the null hypothesis in finite samples. However, as we mention above, as the sample size approaches infinity, the estimate converges to the true parameter value (assuming the estimator is consistent) and its variance approaches 0. Thus, the power will approach 1; the estimator will reject the false null hypothesis every time. Lastly, the power of an estimator depends on the significance level of the test. For example, while an estimator may reject the null hypothesis at the 10% significance level, it may fail to do so at the 5% significance level. While the researcher may choose to report power across all three common significance levels (1%, 5%, 10%), we recommend that the researcher at the very least report power at the 5% significance level since it is the conventional level of hypothesis testing.

Appendix C. Other Performance Statistics and the Full Pattern of Reporting

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" and/or "simulation."[1] As Figure 1 shows, a search for these terms identified a total of 540 articles. From this initial set of articles, we identified 71 in which the results from a Monte Carlo simulation were reported in the published article.[2] As detailed in Section 4 of the manuscript, we found that measures of bias are most commonly reported, followed by MSE/RMSE. In descending order of frequency, we found that coverage, standard deviation, overconfidence, and power are least reported.

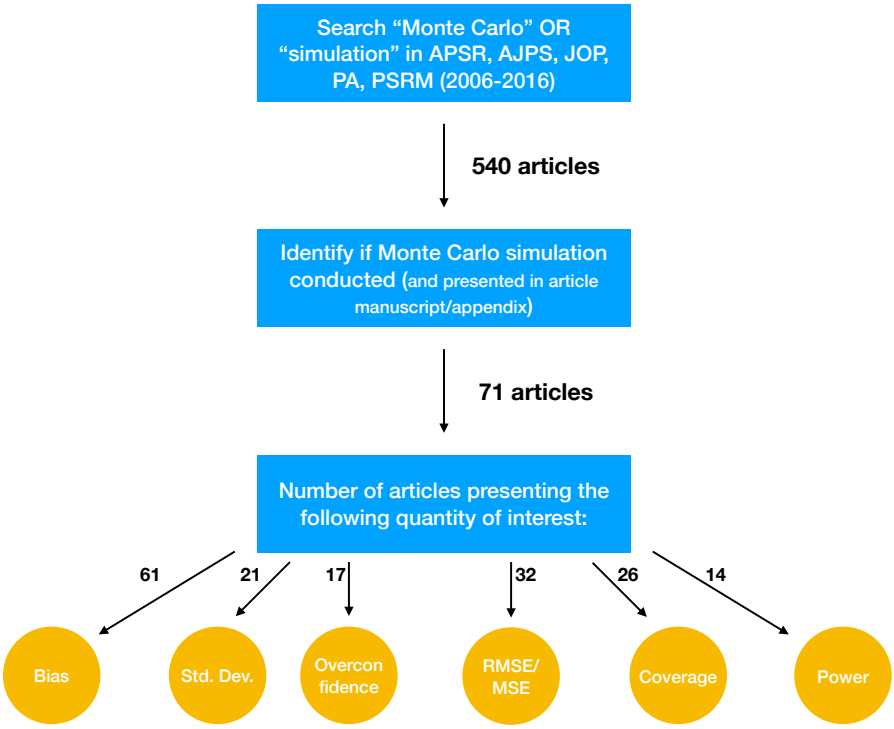

Figure 1: Summary of our literature coding

In Table 1 in the main text we presented patterns of performance statistics, reporting only the patterns that were used by more than 4.2% of studies for brevity. In Table 2 in this document we show the full pattern of reporting.

[1]Since publication of *Political Science Research and Methods* began in 2013, we coded 2013-2016 for that journal.

[2]We coded all Monte Carlo simulations that were presented as a part of published papers and in appendices that appeared as a part of the volume in which they were published. We did not code the use of Monte Carlo simulations that only appeared in supplemental materials made available online.

Performance Statistic							
Bias	RMSE or MSE	Coverage/ Type 1	SD	Overconfidence	Power/ Type 2	Pattern Percentage	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
		C			P	2.8	average error and its sources; the sources of inference problems
B		C	S	O		2.8	average error and power
B	R		S			2.8	inference problems and one source
B	R		S	O		2.8	inference problems
B	R	C	S			2.8	power and one source of inference problems
					P	1.4	average error and its sources; coverage; sources of inference problems
	R	C			P	1.4	sources of both average error and inference problems
B					P	1.4	average error and one source; coverage; two sources of inference problems
B				O	P	1.4	average error and one source; coverage; two sources of inference problems
B	R				P	1.4	one source of average error; coverage; two sources of inference problems
B	R		S		P	1.4	coverage; one source of inference problems
B	R	C		O		1.4	one source of average error; power; one source of inference problems
B	R	C			P	1.4	one source of average error; two sources of inference problems
B	R	C	S	O	P	1.4	
85.9	45.1	36.6	29.6	23.9	19.7	Overall use	

Table 2: Full pattern of reporting performance statistics in major Political Science journals

Notes: The letters in each row indicate that that particular performance statistic was reported for studies referenced in that row. B–bias, R–RMSE, C–coverage, S–SD, O–overconfidence, P–power.

In addition to examining the use of our recommended performance statistics as shown in Table 2, we also assessed the degree to which other performance statistics were used. We found two other commonly used performance statistics: percentile range and density plots.

- **Percentile Range:** Percentile range constitutes the lower and upper bounds of a quantity of interest for a given confidence level. For example, the 95% percentile range of a parameter estimate is the 2.5% and 97.5% percentiles of the distribution of estimates from repeated sampling. In other words, percentile range is the confidence interval for the quantity of interest. It is an alternative to SD and is more appropriate when the sampling distribution is asymmetric. 18.31% of the surveyed articles that conducted Monte Carlo experiments reported percentile range. Percentile range can also be calculated for a performance statistic—for example, the confidence intervals around the bias in the parameter estimate.[3]

- **Density Plots:** 23.94% of the surveyed articles that conducted Monte Carlo experiments presented density plots of the estimates of their quantities of interest from repeated sampling. If properly labelled, these plots can provide information about the bias and efficiency (SD) for the quantity of interest.

[3]We also note that Monte Carlo statistics themselves have uncertainty surrounding the presented quantity of interest (Boos and Osborne, 2015). This is known as "Monte Carlo error", which is the "standard deviation of the estimated quantity over repeated simulation studies (Gasparini, 2018, p. 1). Although rarely presented in studies, they can be easily calculated for the quantities of interest we discuss in both `R` (Gasparini, 2018) and Stata (White, 2010).

Appendix D. Clark and Linzer (2015) Standard Case Replication

In the main text, we replicated Clark and Linzer's (2015) analysis for the sluggish case—i.e., when the within-unit variation of the predictor is 0.2. In this section, we replicate Clark and Linzer (2015) and calculate our recommended performance statistics for the standard case—when the within-unit variation of the predictor is 1.

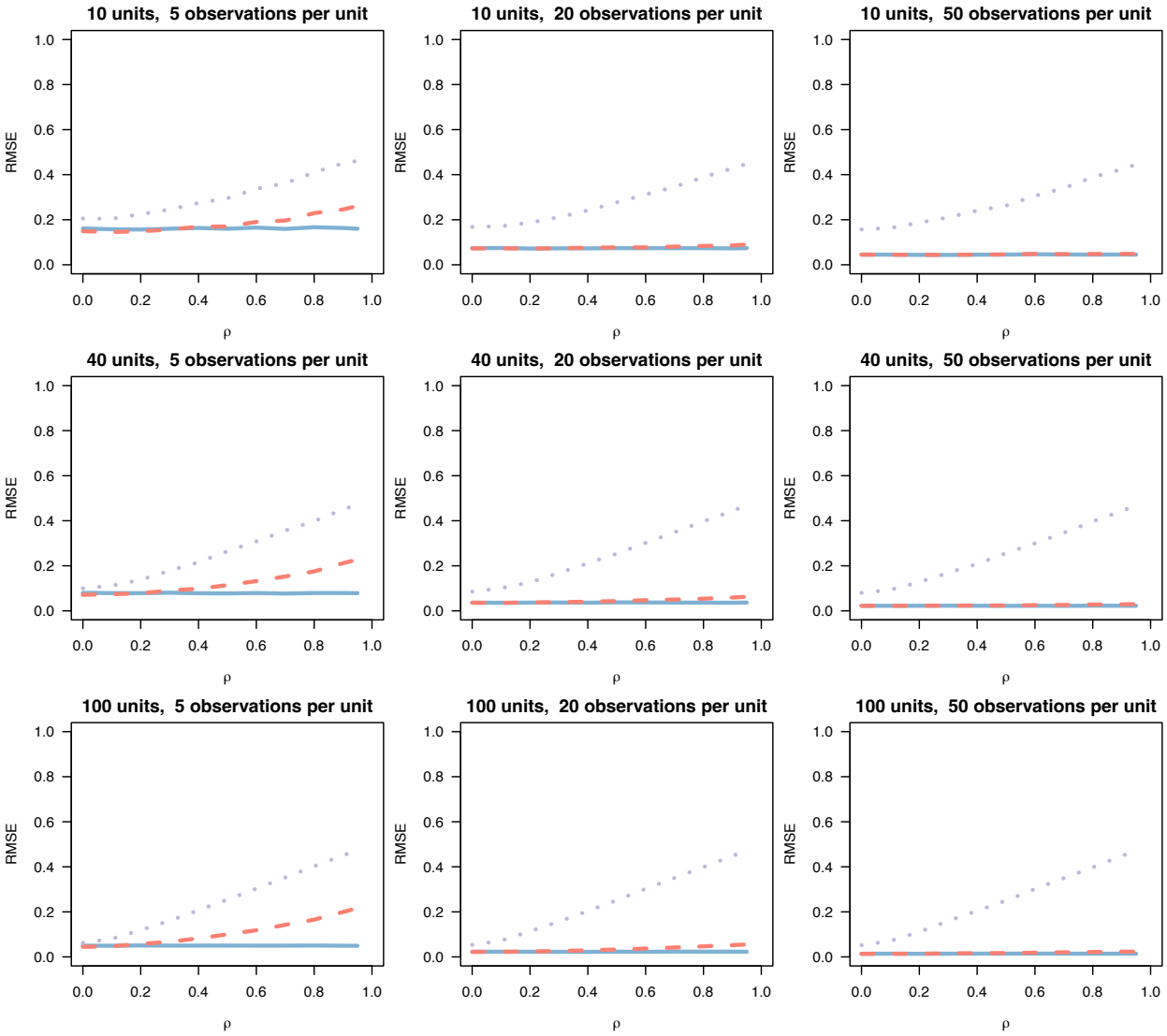

Figure 2: RMSE of $\hat{\beta}$, Clark and Linzer's standard case

Notes: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

To evaluate the estimators' performance in terms of point estimates and inferences, we first calculated RMSE, coverage, and power. In figure 2, we present the RMSEs for the pooled-OLS, fixed effects, and random effects estimators for different combinations of the

number of units (J) and the number of within-unit observations (n). Overall, similar to Clark and Linzer (2015), we find that for all combinations of J and n, the fixed effects estimator has the lowest RMSE. More specifically, the fixed effects and random effects estimators perform similarly in scenarios in which the correlation between the unit effects and predictor (ρ) is low and/or $n \geq 20$. In all other scenarios, the fixed effects estimator performs the best.

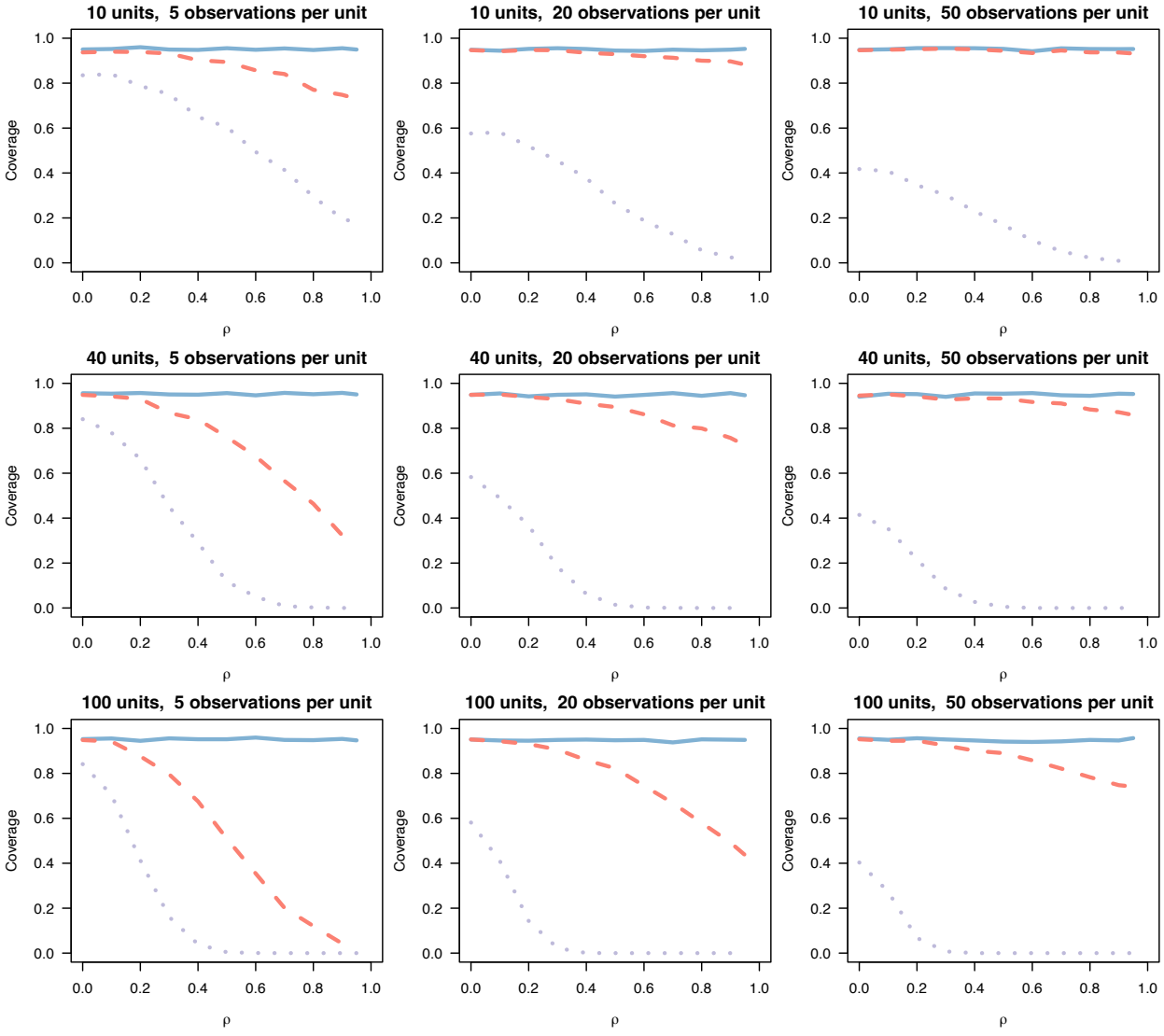

Figure 3: Coverage of $\hat{\beta}$, Clark and Linzer's standard case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

In terms of coverage, as shown in Figure 3, only for low values of ρ do the fixed and random effects estimator perform similarly under most combinations of J and n, and the pooled-OLS estimator always performs worse. At higher values of ρ, the coverage of the fixed effects estimator is always better than that of the random effects estimator. Thus,

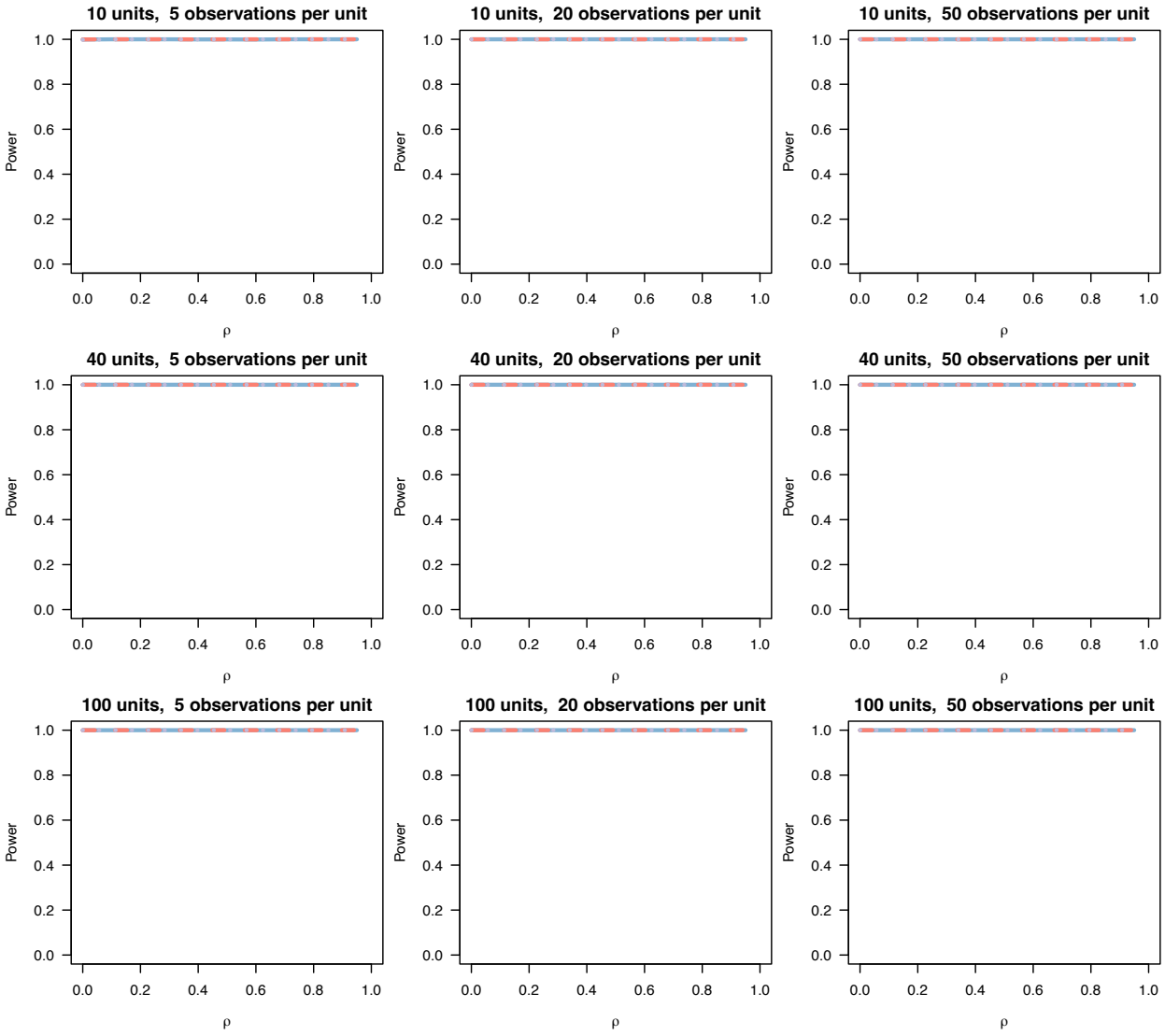

Figure 4: Power of $\hat{\beta}$, Clark and Linzer's standard case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

the fixed effects estimator performs best overall. In terms of power, the pooled-OLS, fixed effects, and random effects estimators, all have a power of 1 under all combinations of J, n, and ρ (Figure 4). They always reject the false null hypothesis that the predictor has no effect on the outcome.

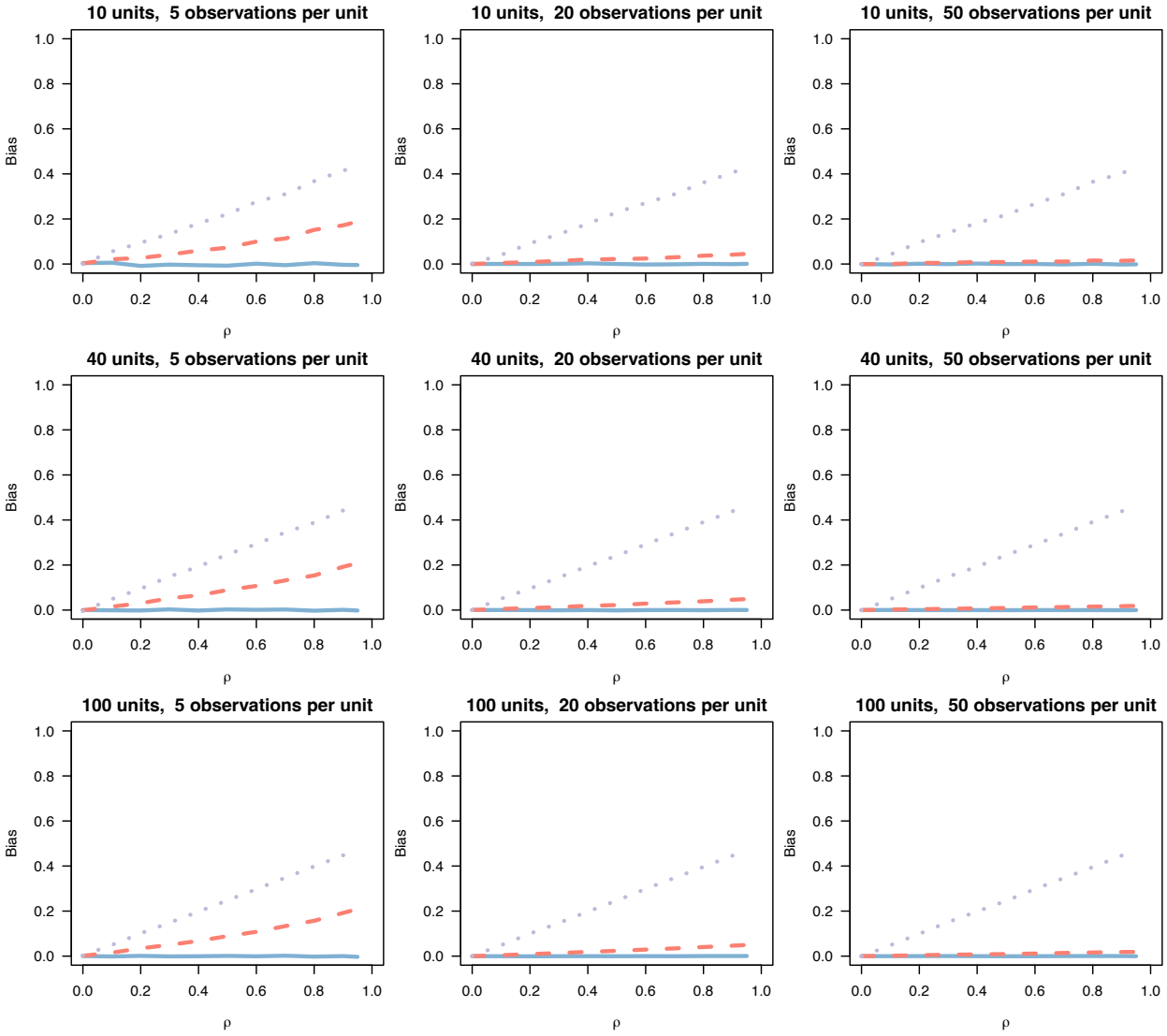

Figure 5: Bias of $\hat{\beta}$, Clark and Linzer's standard case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

To diagnose the estimators' performances in terms of RMSE, coverage, and power, we calculate bias, SD, and overconfidence. Looking at both the bias and SD of the estimators in Figures 5 and 6 provides information about the individual contributions of bias and efficiency to the average error (RMSE). We find that the reason behind the higher RMSE of the random effects estimator at $n \leq 20$ and high ρ is mainly because of its bias. The random effects estimator is only slightly more efficient (lower SD) than the fixed effects

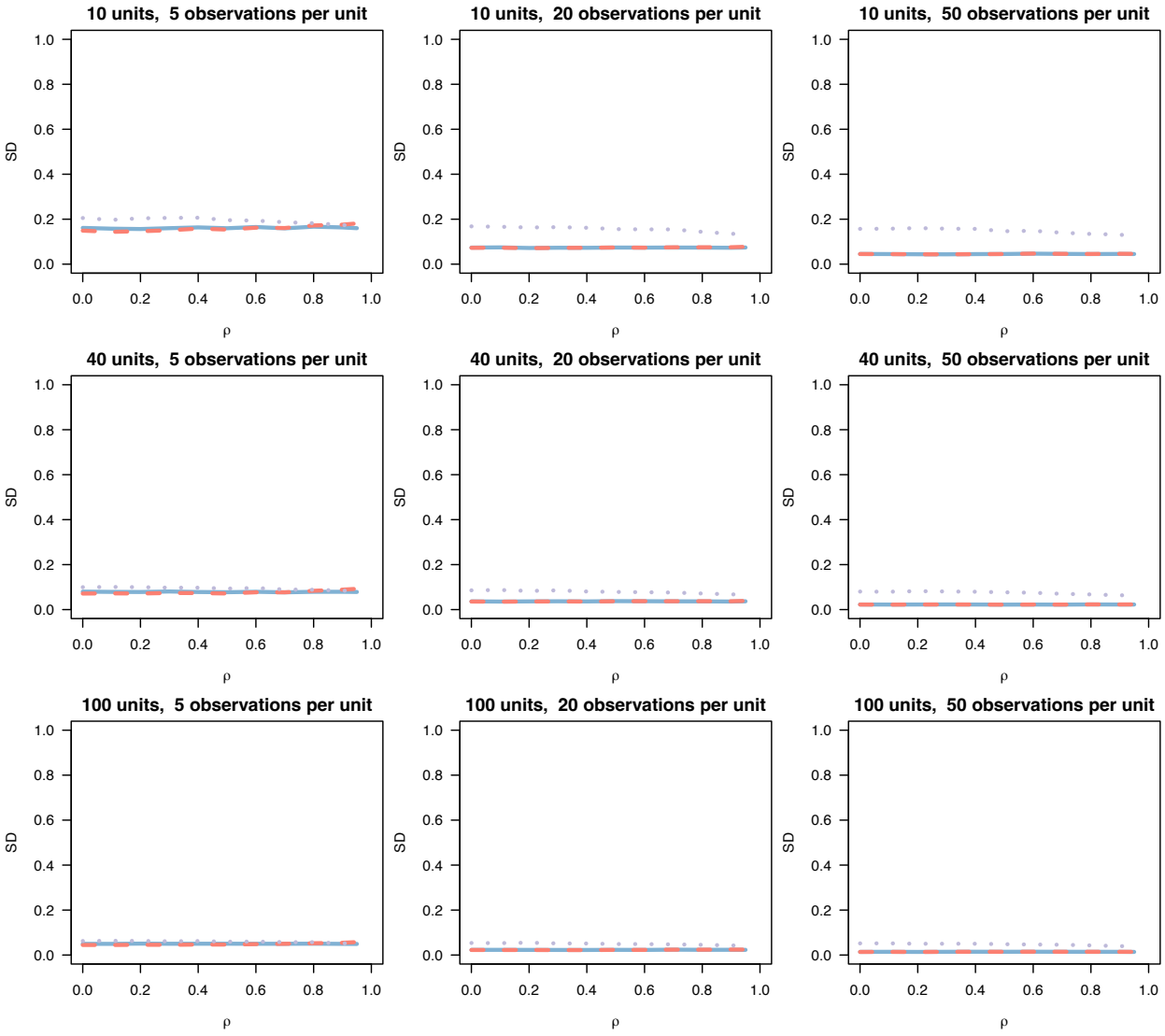

Figure 6: Standard deviation of $\hat{\beta}$, Clark and Linzer's standard case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model, the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

estimator at $J \leq 40$, $n = 5$, and small ρ. Thus, although the fixed estimator is theorized to be inefficient and the random effects estimator is theorized to be biased, we provide a more nuanced conclusion than Clark and Linzer by demonstrating that in the standard case, the fixed estimator is as efficient as the random effects estimator and that the low RMSE of the random effects estimator is due to bias.

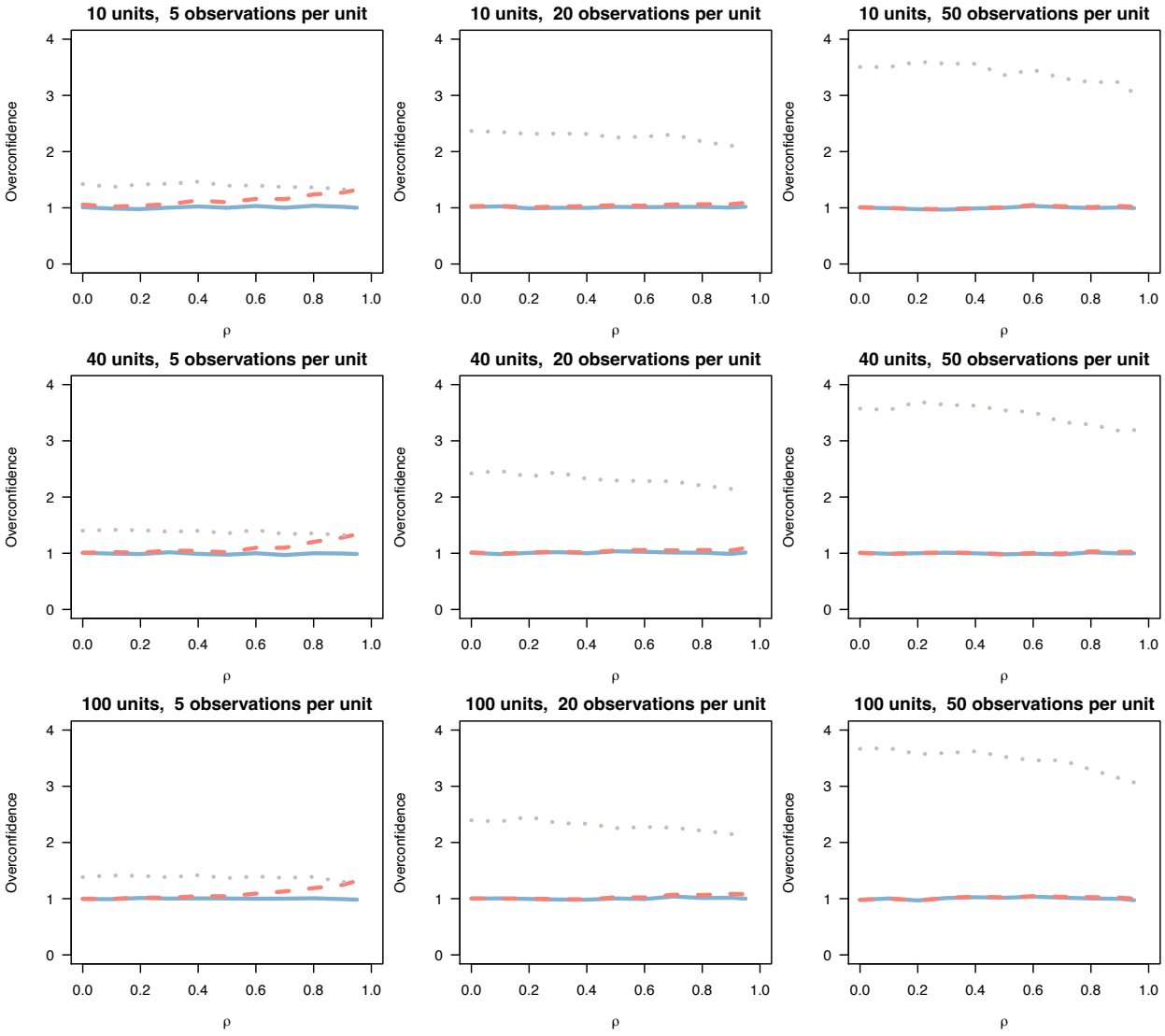

Figure 7: Overconfidence of $\hat{\beta}$, Clark and Linzer's standard case

Note: Solid line = OLS-fixed effects, dashed line = FGLS-random effects, dotted line = OLS-pooled model the horizontal axis in each plot is the value of correlation between $\bar{x}_j$ and α_j (ρ).

In addition to bias and SD, calculating overconfidence provides a more comprehensive understanding of the coverage and power of the three estimators. In figure 7, we present the results for overconfidence. For all combinations of J and n, the fixed effects estimator recovers close-to-accurate estimates of the standard deviation, meaning that its high coverage is because of its unbiasedness, its low SD, and its ability to recover accurate

standard errors. Pooled-OLS is always overconfident and underestimates the standard deviation of the effect of the predictor on the outcome (β), implying that its poor coverage is largely due to a combination of its bias and overconfidence. The random effects estimator performs worse than the fixed effects estimator when $n = 5$ and $J < 100$, and also at high levels of ρ when $n < 50$ due to a combination of its bias and overconfidence. When $n = 50$, its lower than expected coverage is due to its slight bias. The high power of the random effects estimator (when $n = 50$) and the fixed effects estimator (across the board) results from their unbiasedness, efficiency, and accurate standard errors. The random effects estimator's bias only slightly contributes to its power when $n = 5$ or when $n = 20$ and $\rho \geq 0.5$. The pooled-OLS estimator's power is always a function of its biasedness and underestimated/overconfident standard errors.

Based on the results of Figure 2, Clark and Linzer (2015) (p. 404) concluded that "Researchers should feel secure using either fixed- or random-effects models under standard conditions, as dictated by the practical and theoretical aspects of a given application." However, when calculating our recommended performance statistics, our results demonstrate that even under standard conditions, the fixed effects estimator should almost always be preferred. While both estimators often have similar RMSE and power, the fixed effects estimator performs significantly better than the random effects estimator in terms of bias, overconfidence, and coverage, and therefore Type 1 errors.

Appendix E. Wilkins (2018) Replication

Wilkins (2018) weighs in on the highly-cited debate between Achen (2000) and Keele and Kelly (2006) about the use of lagged dependent variables (which Wilkins abbreviates as "LGDVs") in time series models. He starts with the same DGP as the other authors:

$$Y_t = \alpha Y_{t-1} + \beta X_t + u_t \tag{4}$$

$$X_t = \rho X_{t-1} + e_{1t} \tag{5}$$

$$u_t = \phi u_{t-1} + e_{2t} \tag{6}$$

where e_{1t} and e_{2t} are independent and identically distributed stochastic terms that conform to the usual OLS assumptions about error terms. Wilkins points out that researchers have typically estimated Equation 4. He shows that a transformation of this model is equivalent to an Autoregressive Distributed Lag (ADL) of order (2,1) (p. 395),[4]

$$\text{EQ4: } Y_t = (\alpha + \phi)Y_{t-1} + (-\alpha\phi)Y_{t-2} + \beta X_t + (-\beta\phi)X_{t-1} + e_{2t} \tag{7}$$

which Wilkins calls Equation 4 (EQ4). He argues that excluding Y_{t-2} and X_{t-1} from the model when estimating Equation 4 results in omitted variable bias because Y_{t-1} and X_t will be correlated with u_t, which leads to a biased coefficient on X_t. Wilkins uses Monte Carlo analyses to compare the performance of EQ4 with two typically estimated time series models which he labels as "LGDV" and "LGDV2" specified as

$$\text{LGDV: } Y_t = \alpha Y_{t-1} + \beta X_t + u_t$$

$$\text{LGDV2: } Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \beta X_t + u_t$$

We replicate Wilkins' Monte Carlo experiments which he displays in Figures 1(c) and 1(d) on p. 398.[5] Following Wilkins, we set the parameters in the DGP (Equations 4 to 6) such that $\beta = 0.5$, $\rho = 0.95$, and $\alpha = 0.75$, while allowing ϕ, the coefficient of autocorrelation between u_t and u_{t-1} to vary between 0 and 0.5. For each of the resulting scenarios, we also perform 1000 simulations for time series with 100 observations. In addition to replicating the performance of these estimators in terms of bias and RMSE for β, we extend Wilkins' analysis by assessing these estimators' performance in terms of SD, overconfidence, power, and coverage.[6]

The results from this replication and extension of Wilkins' experiment are displayed in Figure 8. In evaluating the performances of the three models in terms of the short-run effect (β), based on the RMSE results, we find that the LGDV and LGDV2 models perform

[4]Wilkins uses standard time series notation for the order of an ADL model in which the first number is the number of lags of the dependent variable and the second number is the number of lags in the independent variable.

[5]We omit reporting results from the "REG" model—$Y_t = \beta X_t + u_t$—because it performs drastically worse than the EQ4, LGDV, and LGDV2 models under all scenarios for the given set of parameter values.

[6]Wilkins reports percent bias which is calculated as $\frac{E(\hat{\beta})-\beta}{\beta} \times 100$. This measure is different from the more common measure of bias that we present in Equation 1 in the manuscript, in that it is scaled relative to β.

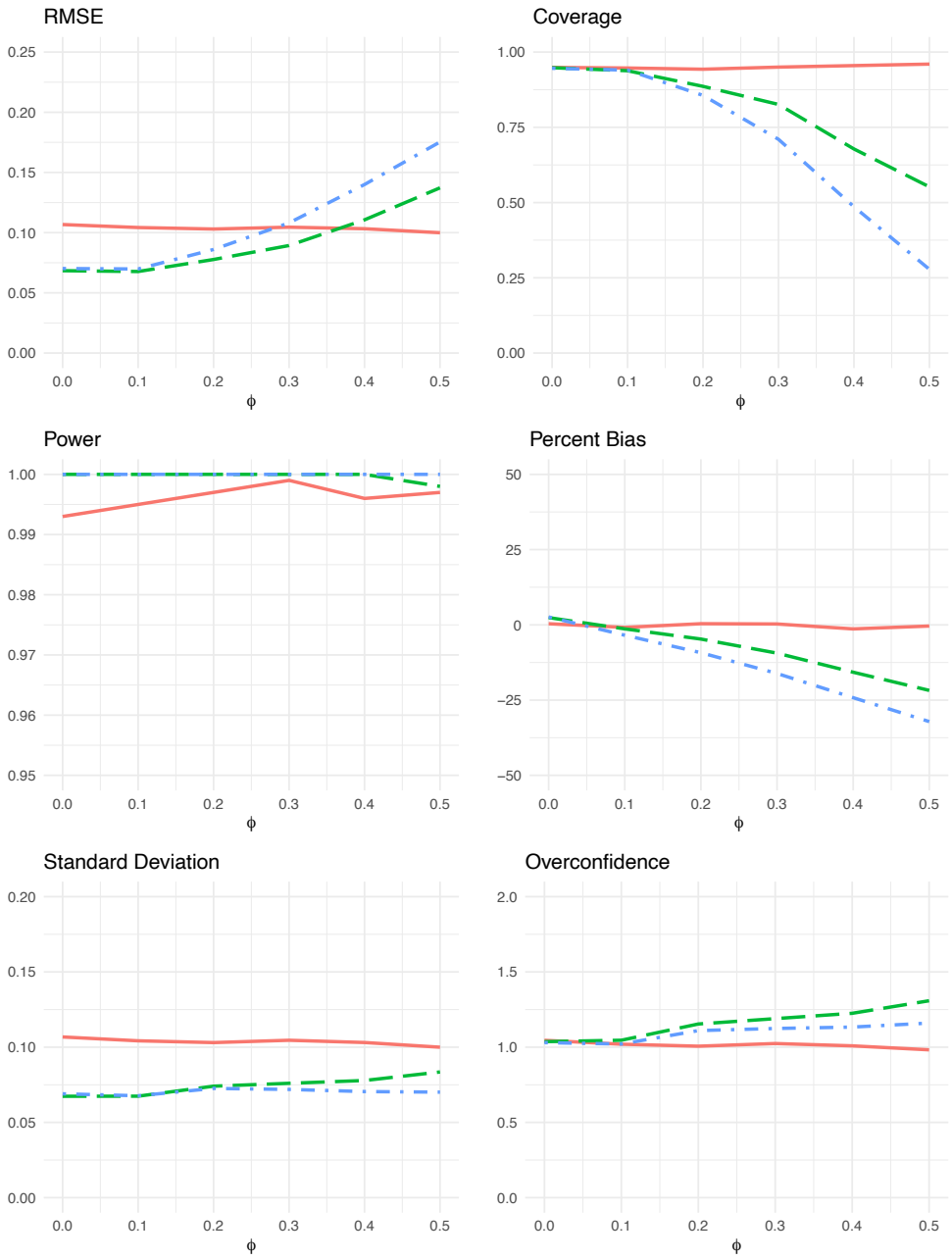

Figure 8: MC performance statistics of Wilkins' Figures 1c and 1d

Note: Solid line = EQ4, dashed line = LGDV, dash/dot line = LGDV2

best in terms of RMSE at low levels of serial autocorrelation ($\phi \leq 0.3$) and EQ4 performs best when $\phi \geq 0.4$. The results for coverage, in the top-right panel of Figure 8 provide additional evidence in favor of the EQ4 estimator for β. Only up to $\phi = 0.1$ are the coverage statistics of the estimators similar across models.[7] For values of $\phi > 0.1$, the confidence intervals of the LGDV and LGDV2 models fail to encompass the true parameter, β, 95% of the time, and this worsens for increasing values of ϕ. This implies the LGDV and LGDV2 models are prone to Type 1 errors. In contrast, the coverage statistic for β from estimating EQ4 is around 0.95 and rises slightly as ϕ increases. But overwhelmingly, the coverage is flat across values of ϕ. Thus, the confidence intervals of $\hat{\beta}$ from EQ4, encompass the true parameter a little over 95% of the time. From the middle-left panel in Figure 8, it is apparent that the estimators of β in all three models have very high power, between 0.99 and 1, across all levels of serial autocorrelation.

In diagnosing the resulting RMSE values, we find that the variation in RMSE between the three models is due to a bias-SD tradeoff. Across all simulated values of ϕ, EQ4 is the least efficient but unbiased and the bias in the LGDV and LGDV2 models are not offset by their efficiency gains at higher levels of ϕ, which result in higher RMSE values than EQ4. The good coverage of EQ 4—despite having the highest SD—is becuase it is unbiased and recovers accurate standard errors (overconfidence of 1) across all values of ϕ. When $\phi > 0.1$, the LGDV and LGDV2 models are biased and overconfident, thus resulting in poor coverage or increased Type 1 errors. The high power of the LGDV and LGDV2 models are a result of their low standard deviation and underestimated standard errors, which offset their (negative) biasedness towards 0.

By examining only percent bias and RMSE—the performance statistics reported by Wilkins—one would conclude in favor of the LGDV model when $\phi \leq 0.3$ because of its low RMSE and in favor of EQ4 when $\phi > 0.3$. However, when looking at coverage, EQ4 performs best as ϕ increases. A similar conclusion holds true in terms of the accuracy of the models' standard errors. Although each performance measure provides important information, because of EQ4's ability to to cover the true effect of the independent variable on the dependent variable (low Type 1 errors), we conclude in favor of EQ4. Such a conclusion would not have been reached if one were to merely look at the performance statistics—percent bias and RMSE—reported by Wilkins. Thus, even for low values of ϕ (e.g., $\phi \leq 0.3$), EQ4 is the best performing model.

[7] The coverage statistics reported here are for 95% confidence levels.

Appendix F. Hanmer and Kalkan (2013) Replication

Hanmer and Kalkan (2013) use Monte Carlo analyses to help make their case for calculating marginal effects from binary-outcome models using an average marginal effect (AME) approach instead of a marginal effects at means (MEM) approach. In the MEM approach, researchers construct a hypothetical observation with mean values for all independent variables and then calculate the impact of a shift in a single independent variable, x_k, on that hypothetical observation such that

$$\text{MEM} = f(\bar{\boldsymbol{x}}\beta)\beta_k.$$

In contrast, the AME approach advocated by Hanmer and Kalkan involves calculating the marginal effect of a shift in a single independent variable, x_k, for each observation in the sample at its observed values for all independent variable values, $\boldsymbol{x}_i\beta$, and then averaging over these marginal effects,

$$\text{AME} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i\beta)\beta_k.$$

To demonstrate the relative utility of these two approaches, Hanmer and Kalkan generate binary-outcome data using the following DGP:

$$y^* = 2 + -1x_1 + 1x_2 + 0.5x_3 + e$$

where, $y = 1$ if $y^* > 0$ and $y = 0$ if $y^* \leq 0$. x_1 takes on the values 1, 2, or 3 and is created from a uniform distribution that is divided into three equally probable categories. x_2 and x_3 are drawn from standard normal distributions and can have a 0, 0.5, or 0.8 correlation with each other. They then estimate four probit models: 1) y is regressed on x_1, x_2, and x_3 in the True Model; 2) y is regressed only on x_2 and x_3 in Model 1; 3) y is regressed only on x_1 and x_3 in Model 2; and, 4) y is regressed only on x_1 and x_2 in Model 3. They calculate bias in both the MEM and AME for each of the latter three models by subtracting these estimated effects from their values in the true model. Across the board, they find that the bias of the estimated MEM marginal effects are greater than those of the estimated AME effects.

We expand their analysis to include RMSE, coverage, and power, to evaluate the estimators' performances, and standard deviation and overconfidence to diagnose those performances. We replicate Panel A of Table 1 (p. 274) when the correlations between x_2 and x_3 are 0, 0.5, and 0.8. Following the lead of Hanmer and Kalkan (2013), we also conduct 1000 simulations for a sample size equal to 1000.

Our results for Model 1—when y is regressed only on x_2 and x_3—are presented in Table 3. By analyzing only the bias in marginal effects across all models, Hanmer and Kalkan find that the AME approach recovers unbiased estimates of the coefficients of x_2 and x_3, when x_1 is omitted. This, however, leaves the reader uncertain as to whether the AME approach has lower RMSE and higher coverage and power, and whether it is also more efficient and recovers accurate standard errors.

	Model 1: *Excludes* x_1 $\text{Cor}(x_2, x_3) = 0$		Model 1: *Excludes* x_1 $\text{Cor}(x_2, x_3) = 0.5$		Model 1: *Excludes* x_1 $\text{Cor}(x_2, x_3) = 0.8$	
	MEM	AME	MEM	AME	MEM	AME
			True Marginal Effects Values			
x_2	0.401	0.231	0.401	0.214	0.401	0.206
x_3	0.201	0.116	0.201	0.107	0.201	0.103
			Root Mean Squared Error			
x_2	0.099	0.007	0.098	0.008	0.099	0.011
x_3	0.050	0.007	0.050	0.008	0.051	0.011
			Coverage			
x_2	0	0.993	0.002	0.997	0.084	0.998
x_3	0.157	0.999	0.376	0.999	0.778	0.999
			Power			
x_2	1	1	1	1	1	1
x_3	1	1	1	1	1	1
			Bias			
x_2	−0.098	0.000	−0.097	0.000	−0.096	0.000
x_3	−0.049	0.000	−0.048	0.000	−0.048	0.000
			Standard Deviation			
x_2	0.020	0.011	0.024	0.013	0.033	0.020
x_3	0.019	0.013	0.023	0.015	0.032	0.021
			Overconfidence			
x_2	0.963	0.995	0.977	1.017	0.981	0.997
x_3	1.027	1.038	1.033	1.037	1.012	1.012

Table 3: MC Performance Statistics of Hanmer and Kalkan's Table 1, Panel A.

Note: Coverage probabilities are calculated for the 95% confidence level.

Table 3 demonstrates that that the RMSE values of the AME approach are substantially smaller than those from the MEM approach. Under all reported scenarios of omitted variables and various Cor(x_2, x_3), the MEM approach has poor coverage. That is, the estimated marginal effect only rarely encompasses the true marginal effect at the means, increasing Type 1 errors. The AME approach, on the other hand, has coverage values greater than 0.95 for a 95% constructed confidence interval. Both marginal effects approaches have a power of 1 for any level of correlation between x_2 and x_3. In other words, both approaches perform equally well with respect to Type 2 errors.

To diagnose the performances of these approaches on RMSE, coverage, and power, we calculate bias, SD, and overconfidence. From the bias and standard deviation values, we can infer that the MEM approach's high RMSE is due to both its bias and inefficiency, and that the AME approach's (lower) RMSE values are due to its (lower) standard deviation values and the fact that it is unbiased. From the overconfidence results, we see that, although the AME approach is more efficient than the MEM approach, both approaches perform similarly in recovering accurate standard errors: they either slightly underestimate (overconfident) or overestimate (underconfident) the standard errors. This tells us that the poor coverage or higher Type 1 errors exhibited by the MEM approach are due to its bias and SD, and that bias and overconfidence contribute to the MEM approach's high power.

Although our replication of Model 1, Panel A, Table 1 (Hanmer and Kalkan, 2013) reaches a similar conclusion as that of the authors—the AME approach is superior to the MEM approach—our approach demonstrates that the superiority of the AME approach across the scenarios examined is comprehensive by providing two new findings. First, the AME approach has fewer Type 1 errors because it is both less biased and more efficient than the MEM approach. Second, the AME approach has a lower RMSE because it is both less biased and more efficient than the MEM approach. We also find that the AME and MEM approaches perform equally well with respect to Type 2 errors under our test procedure (power). Our replication and extension demonstrates the robustness of the AME approach by extending the analyses from merely reporting bias, to finding that the AME approach performs substantially better in terms of coverage, RMSE, and SD. However, this is not true for power, a performance statistic in which the AME approach is as robust as the MEM approach.

It is important to note that in Model 1—when x_1 is excluded—x_1 is not correlated with x_2 or x_3, and thus the consequences of omitted variable bias should ideally not appear in the marginal effects calculations for x_2 and x_3. Despite this, the MEM approach performs poorly. In Tables 4 and 5, we replicate the rest of Panel A, Table 1 in Hanmer and Kalkan (2013) by calculating our recommended performance statistics, and thus probe the robustness of the AME approach when a relevant variable (x_2 or x_3) is omitted from the model. As per Hanmer and Kalkan (2013), the correlation between x_2 and x_3 is either 0, 0.5, or 0.8. In evaluating the approaches based on these simulations, we find that, for any level of correlation, when either x_2 or x_3 is excluded, the AME approach has a lower RMSE than the MEM approach. In terms of coverage, the AME approach performs well, and better than the MEM approach, when the correlation between x_2 and x_3 is 0. For any non-zero correlation between x_2 and x_3, both approaches have a coverage of zero. Further, both approaches have a power of 1 for any level of correlation between x_2 and x_3.

	Model 2: *Excludes* x_2 $\text{Cor}(x_2, x_3) = 0$		Model 2: *Excludes* x_2 $\text{Cor}(x_2, x_3) = 0.5$		Model 2: *Excludes* x_2 $\text{Cor}(x_2, x_3) = 0.8$	
	MEM	AME	MEM	AME	MEM	AME
			True Marginal Effects Values			
x_2	0.401	0.231	0.401	0.214	0.401	0.206
x_3	0.201	0.116	0.201	0.107	0.201	0.103
			Root Mean Squared Error			
x_2	–	–	–	–	–	–
x_3	0.061	0.008	0.102	0.107	0.246	0.165
			Coverage			
x_2	–	–	–	–	–	–
x_3	0.029	0.999	0	0	0	0
			Power			
x_2	–	–	–	–	–	–
x_3 1	1	1	1	1	1	1
			Bias			
x_2	–	–	–	–	–	–
x_3	−0.060	0.000	0.101	0.106	0.245	0.164
			Standard Deviation			
x_2	–	–	–	–	–	–
x_3	0.018	0.014	0.021	0.011	0.027	0.010
			Overconfidence			
x_2	–	–	–	–	–	–
x_3	1.012	1.026	0.983	1.007	0.991	1.134

Table 4: MC Performance Statistics of Hanmer and Kalkan's Table 1, Panel A, when excluding x_2.

Note: Coverage probabilities are calculated for the 95% confidence level.

Thus the consequences of omitted variable bias manifest clearly when examining coverage, which demonstrates that both approaches perform poorly in terms of Type 1 errors. In diagnosing the performances of the two approaches, we find that he lower RMSE of the AME approach is because it is less biased and more efficient than the MEM approach. For the AME approach, despite being more efficient than the MEM approach, its poor coverage is also a consequence of its overconfidence, especially when x_3 is excluded from the model (Table 5). Although the authors claim that the AME approach is more robust under conditions of omitted variable bias (p. 273), we reach a more nuanced conclusion. Overall, while the AME approach is still the preferred approach, we find that the AME approach will always reject the null hypothesis when it is true. In other words, in the face of omitted variable bias as created in these scenarios, both the AME and MEM approaches will always incorrectly reject the true null hypothesis. To conclude, the consequences of omitted variable bias are manifested as low coverage, and thus increased Type 1 errors for both the AME and MEM approaches. However, the AME still does better in terms of bias, efficiency, and RMSE. The AME and MEM approaches perform similarly in terms of power.

	Model 3: *Excludes* x_3 $\text{Cor}(x_2, x_3) = 0$		Model 3: *Excludes* x_3 $\text{Cor}(x_2, x_3) = 0.5$		Model 3: *Excludes* x_3 $\text{Cor}(x_2, x_3) = 0.8$	
	MEM	AME	MEM	AME	MEM	AME
			True Marginal Effects Values			
x_2	0.401	0.231	0.401	0.214	0.401	0.206
x_3	0.201	0.116	0.201	0.107	0.201	0.103
			Root Mean Squared Error			
x_2	0.045	0.005	0.059	0.054	0.138	0.084
x_3	–	–	–	–	–	–
			Coverage			
x_2	0	1	0	0	0	0
x_3	–	–	–	–	–	–
			Power			
x_2	1	1	1	1	1	1
x_3	–	–	–	–	–	–
			Bias			
x_2	−0.043	0.000	0.058	0.054	0.136	0.082
x_3	–	–	–	–	–	–
			Standard Deviation			
x_2	0.024	0.010	0.029	0.010	0.033	0.010
x_3	–	–	–	–	–	–
			Overconfidence			
x_2	1.008	1.053	1.011	1.196	1.012	1.314
x_3	–	–	–	–	–	–

Table 5: MC Performance Statistics of Hanmer and Kalkan's Table 1, Panel A, when excluding x_3.

Note: Coverage probabilities are calculated for the 95% confidence level.

References

Achen, Christopher H. 2000. Why lagged dependent variables can suppress the explanatory power of other independent variables. In *annual meeting of the political methodology section of the American political science association, UCLA*. Vol. 20 pp. 07–2000.

Boos, Dennis D and Jason A Osborne. 2015. "Assessing variability of complex descriptive statistics in monte carlo studies using resampling methods." *International Statistical Review* 83(2):228–238.

Clark, Tom S and Drew A Linzer. 2015. "Should I use fixed or random effects?" *Political Science Research and Methods* 3(02):399–408.

Gasparini, Alessandro. 2018. "rsimsum: Summarise results from Monte Carlo simulation studies." *Journal of Open Source Software* 3(26):739.

Greene, William H. 2017. *Econometric Analysis*. Eight ed. New York, NY: Pearson.

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):263–277.

Keele, Luke and Nathan J Kelly. 2006. "Dynamic models for dynamic theories: The ins and outs of lagged dependent variables." *Political analysis* pp. 186–205.

White, Ian R. 2010. "simsum: Analyses of simulation studies including Monte Carlo error." *The Stata Journal* 10(3):369–385.

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.