Advance Access publication January 4, 2007

Political Analysis (2007) 15:101–123
doi:10.1093/pan/mpl012

A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications

Sven E. Wilson
Department of Political Science, Brigham Young University, 732 SWKT, Provo, UT 84602
e-mail: sven_wilson@byu.edu (corresponding author)

Daniel M. Butler
Department of Political Science, Stanford University, Encina Hall West, Room 100, Stanford, CA 94305
e-mail: daniel_butler@stanford.edu

In 1995, Beck and Katz (B&K) instructed the profession on "What to do (and not to do) with time-series, cross-section data," and almost instantly their prescriptions became the new orthodoxy for practitioners. Our assessment of the intellectual aftermath of this paper, however, does not inspire confidence in the conclusions reached during the past decade. The 195 papers we reviewed show a widespread failure to diagnose and treat common problems of time-series, cross-section (TSCS) data analysis. To show the importance of the consequences of the B&K assumptions, we replicate eight papers in prominent journals and find that simple alternative specifications often lead to drastically different conclusions. Finally, we summarize many of the statistical issues relative to TSCS data and show that there is a lot more to do with TSCS data than many researchers have apparently assumed.

1 Introduction

Roughly two decades ago, Stimson (1985) published an oft-cited review essay in *American Journal of Political Science* on basic methods of approaching panel data. At the beginning of this essay, he warned readers that the statistical issues associated with these types of data sets are "formidable" (p. 914). At the end, the tone was similar: "To deal with the complications of pooled design in detail makes us painfully aware of a plethora of problems," though dealing with these problems is "sometimes worth its price." His message was that

Authors' note: We greatly benefited from the comments of Neil Beck, Richard Butler, Damon Cann, Scott Cooper, Jay Goodliffe, Donald Green, Darren Hawkins, Daniel Nielson, and Michael Thies. Joseph Burton provided excellent research assistance. We also express thanks to the authors who graciously provided data for this study and subjected themselves to our critique: Michael Campenni, Gary Cox, M. V. Hood, Quentin Kidd, David Lanoue, Karl Moene, Irwin Morris, Jeffrey Pickering, Steven Poe, Gary Reich, Frances Rosenbluth, Steven Saideman, Samuel Stanton, Neal Tate, Michael Thies, Michael Wallerstein, and Nikolas Zahariadis. These data were provided to us either directly or through a publicly available Web site. In either case, the authors' cooperation is commendable and appreciated. Supplementary materials for this article are available on the *Political Analysis* Web site.

although a variety of estimation techniques were available, the estimator of choice depended fundamentally on the research design and the "situational context." (p. 945)

As far as research in political science is concerned, a significant shortcoming of the large literature on panel data methods is that it was developed almost exclusively for econometric study of data sets where the number of units (N) dominates the number of time periods (T) and where N is large enough to rely on the asymptotic properties of the estimator. In 1995, Nathaniel Beck and Jonathan Katz (B&K) wrote a paper entitled "What to do (and not to do) with time-series, cross-section data" in the *American Political Science Review*, in which they argued that many of the data sets used in political science had both a small T and a small N and that generalized least squares (GLS) estimates derived from this type of panel data, which they referred to as time-series, cross-section (TSCS), could not be trusted. They convincingly rejected a commonly used approach to TSCS data, the Parks (1967) method, and cast considerable doubt on many high-profile studies employing this technique.[1] This paper made the hugely important but oft-neglected point that we need to better understand how estimators actually behave in the type of data to which we apply them.

But a concern over misapplying asymptotic estimators was not the only distinguishing feature of this highly influential article. In an early footnote, B&K state that their analysis assumes familiarity with the basics of panel data analysis as laid out by Stimson (1985) and by Hsiao (1986), a leading textbook.[2] But their concluding comments are much different in tone and style from Stimson's. After discrediting the Parks method, they "counterbalance this negative conclusion by providing a simple methodology for analyzing TSCS data." (p. 645) In the concluding paragraphs they articulate this simple method without referring to any of the common specification and estimation issues—Stimson's "plethora of problems"—relevant to *any* panel data set, including TSCS data.

Given the professional stature of the authors, the elite status of the journal, the all-encompassing nature of the paper's title, the straightforward description of the new method, and the very limited attention given to any alternative approaches, it is quite understandable that this paper could be taken by some researchers as highly authoritative and even comprehensive. The question becomes, then, whether researchers using the B&K method carefully considered basic issues associated with TSCS data or whether they single-mindedly followed the method suggested by B&K, with little concern for the formidable challenges that Stimson warned against.

To address this question, we examine the published literature with respect to two basic issues: (i) whether the research considers the question of unit heterogeneity and the assumption that panels can be pooled into one data set with a common intercept and slope coefficient (the pooling assumption is the "critical assumption" [p. 636] according to the B&K method); (ii) whether alternative dynamic structures, either in terms of theoretical arguments or empirical tests, are considered. We confront 195 published papers in political science with these two criteria. Given that we are not really asking a lot from these papers, our results are quite discouraging. We find little discussion or consideration of specification issues and even less sensitivity analysis. In short, a large number of studies do not seem to illustrate an understanding of the basics of panel data methods (as instructed by B&K's footnote). Worst of all, a nontrivial number of studies appear to be nothing more than a blind application of the method of B&K.

[1] Interestingly, though B&K were right to point out the problems with relying on asymptotic estimators, the main weakness in the Park's method turned out to be that it involved the estimation of so many variance parameters that it was only reliable if T is much larger than N, which is a fundamentally different issue than relying on N asymptotics.

[2] The latest edition of this excellent text was published in 2003.

We also replicate the key results of eight papers published in elite journals to see whether our critiques mentioned above actually matter in practice. In other words, do different specifications lead to different results? We test whether the published results are robust to the inclusion of fixed effects (the simplest method to account for unit heterogeneity) and to simple alternative dynamic structures that are different from the lagged dependent variable (LDV) model proposed by B&K. Although some findings hold up under a variety of alternative specifications (a fact that would make the published results stronger had the authors performed and reported some sensitivity analysis), we find a surprising degree of nonrobustness with respect to both unit heterogeneity and dynamic specifications.

Our goal here is not to promote "complicated" (p. 645) estimators warned against by B&K; indeed, we stay solidly in the world of least squares. We certainly endorse B&K's warning that "it is critical that we learn to assess the properties of complicated estimation strategies, and in particular that we study these properties for the types of data actually analyzed . . . ." (p. 654) But we add to their warning a caution not to ignore the well-known problems associated with simple estimators, either. In what follows we examine whether or not basic specification issues associated with TSCS data are being handled with care. We tell a tale that contains a moral both for those who produce methodological advice and for those who consume it. The moral is this: when experts tell us *what* to do, this should not mean there is *nothing else* to do. The problem is not that the B&K papers are full of mistakes (they are not) or that researchers have ignored their advice (they have not); rather, many researchers have followed the prescriptions far too exactly and have overlooked a variety of specification issues, alternative models, appropriate diagnostics, and long-established pitfalls of regression analysis. For the applied researcher, the lesson is to be wary of shortcuts and simple recipes for approaching complex problems. For the methodologist, the lesson is to avoid providing such recipes.

2 Preliminaries

In what follows we consider hierarchical models of the following form:

$$Y_{it} = \beta X_{it} + \alpha_i + u_{it}. \tag{1}$$

The index i refers to the N observational units (or panels), and t indexes the T time periods. The vector of independent variables, X_{it}, may contain lagged values of either X or Y. The α_i term signifies a unit-specific contribution to the dependent variable, and u_{it} is the error term associated with unit i at time t. At this level of generality, we do not place further restrictions on the coefficients or upon the structure of the covariance matrix of the error terms, Ω.[3]

2.1 *B&K in a Nutshell*

The B&K method for TSCS data with continuous dependent variables is captured in two papers (B&K, 1995, 1996), the earlier being more influential. The method consists of three essential components:

1. Pool the data from different units (countries) into one data set and apply ordinary least squares (OLS);

[3]We could also generalize equation (1) further by allowing the slope coefficients to vary across panels—an important type of possible heterogeneity. B&K (2004) have a working paper on this topic, and they conclude that Bayesian approaches to the random coefficients model (RCM) will probably perform best, a conjecture that we agree with. They also note that some RCM models perform very poorly in TSCS samples. Almost none of the published work we have reviewed considers the possibility of the variable coefficients. We certainly concur with B&K that the RCM needs more attention in TSCS studies.

2. Adjust for autocorrelation by either adding an LDV to the model or transforming the data based on an estimate of autocorrelation of the error terms, assumed to be common across panels; and
3. Calculate panel-corrected standard errors (PCSEs).

In terms of component (1), the B&K method is to assume a common intercept for all panels ($a_i = a$). PSCEs are obtained by treating the variance–covariance matrix Ω as an *NTxNT*block diagonal matrix with $\hat{\Sigma}$ along the diagonal, where $\hat{\Sigma}_{i,j} = (\sum_{t=1}^{T} e_{i,t}e_{j,t})/T$.

They then apply the standard formula for OLS residuals with nonspherical error terms:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}.$$

We have no quibble with the use of PCSEs. They embody a reasonable away to account for nonspherical errors within the OLS context. However, they do not account for more fundamental assumptions about the estimating equations themselves, such as whether the common intercept assumption is valid or whether lagged independent or dependent variables should be included in the specification. It is to these issues that we now briefly turn.

2.2 *Heterogeneity*

Unit heterogeneity means that units (countries, states, etc.) differ in ways not explained by observed independent variables. In other words, potentially important local factors are unobservable to the researcher.[4] When researchers use OLS on data pooled from different units, they implicitly assume that unobserved local factors do not exist (meaning that α_i is constant across countries: $\alpha_i = \alpha_j = \alpha$). Figure 1 illustrates the severe consequences that can result from using OLS inappropriately on pooled data. In this example, each of the two countries has data characterized by the simple linear regression model $Y = \alpha_i + \beta X + u$, where u_{it} is a random error term. In all four cases, the error distributions are identical and each country has the same slope coefficient, $\beta = 1$, but the countries have different values for α and different distributions for X. The solid lines in each panel show the regression line estimated by pooled OLS, which can result in overestimating (panel b) or underestimating (panel c) the slope parameter β, including a reversal of sign (panel d). We stress that PCSEs have no effect whatsoever on the bias resulting from using pooled OLS inappropriately.

A variety of estimation techniques exist to estimate equation (1).[5] In our reanalysis of results from the literature (Section 3.2), we use the fixed-effects model[6] (FEM) because it is one of the most common approaches and has the advantage of being unbiased with known small-sample properties as long as the regressors in X_t are exogenous and do not contain an LDV.[7] The FEM can also be a simple diagnostic tool; by comparing the pooled

[4] In the context of the RCM referenced above, heterogeneity could also exist in terms of the effects of observed variables on the dependent variables (i.e., the slope coefficients vary across units).

[5] See, for instance, Baltagi (2002), Wooldridge (2002), Arellano (2003).

[6] Also referred to as the least-squares dummy variable model because it can be estimated simply by adding unit dummies to the OLS regression. All the usual properties of OLS apply.

[7] The most common alternative to the FEM is the random-effects model (REM), which can be estimated with GLS or maximum likelihood estimation. B&K argued that the REM, an estimator that relies on asymptotic properties, is not appropriate for TSCS. We generally concur with this assessment, particularly since the REM requires that the unit effects are uncorrelated with the regressors, a condition that will likely fail in many practical applications. For instance, this assumption is invalid in panels c and d of Fig. 1. We stress, however, that there will be some applications for which the REM is the appropriate choice. If unit effects and the regressors are uncorrelated (which can be examined with a common Hausman [1978] test), the REM is more efficient than FEM and will generally be unbiased (Andrews 1986). However, inference will still rely on asymptotic properties, which may not be appropriate in many applications.

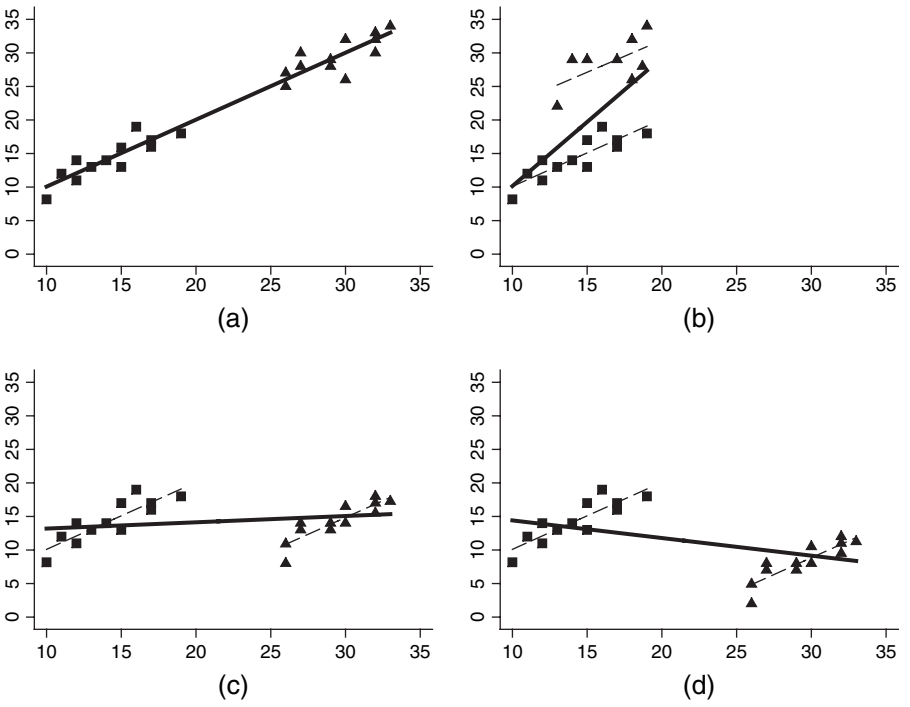

Fig. 1 The thick line in each panel is the estimated slope from the pooled regression. (a) Pooled regression correctly estimates slope; (b) pooled regression overestimates slope; (c) pooled regression underestimates slope; and (d) pooled regression estimates incorrect sign for slope.

OLS and FEM coefficient estimates, a researcher can tell whether unit effects influence the parameters of interests and they can identify which panels (or groups of panels) differ most significantly from the mean. Standard F tests can determine the statistical significance of the unit effects individually, in subsets, or globally.

2.2.1 Sluggish variables

A characteristic (some would say a shortcoming) of the FEM model is that time-invariant variables cannot be included in the model, and slowly moving variables will typically have high standard errors because they will be highly correlated with the fixed effects. We refer here to time-invariant and slowly changing variables as "sluggish." As Beck (2001) writes, "if a variable . . . changes over time, but slowly, the fixed effects will make it hard for such variables to appear either substantively or statistically significant . . .. If an F-test indicates that fixed effects are required, then researchers should make sure they are not losing the explanatory power of slowly changing or stable variables of interest."

Beck warns against "losing" the explanatory power of sluggish variables. In other words, he warns against a type II error—rejecting an effect that really does matter. But the researcher should also be careful of inflating the sample size to produce a type I error—accepting the presence of an effect when it really is not there. Whether a type I error or a type II error is more important depends on the context,[8] but, generally,

[8]Some type II errors can be devastating—such as withholding a safe, lifesaving treatment from patients just because one cannot be 95% sure that the treatment is effective.

conservative scientific inference is concerned with minimizing the probability of a type I error. The cross-sectional variation in the sluggish variables may be an important explanation for the variance in the dependent variable, but if this is true, it will usually show up without relying on inflating the sample size by doing pooled OLS. In short, considering the possibility of unit heterogeneity raises the bar for confirming our theories.

Finally, some might argue that when theory suggests a certain set of explanatory variables, those variables should be included instead of unit effects. After all, should not our models be parsimonious and theoretically motivated? Of course. But to use theory as an argument against the diagnostic value of FEM is to fundamentally misunderstand the role of statistical analysis in theory evaluation. If we *knew* the true model (not that a model is ever really "true") and had all the appropriately measured data, then this would be a valid argument. But absent divination of the true specification, we first use regression analysis to *test* our theories against plausible alternatives. Unit heterogeneity represents the alternative explanation (almost always a plausible one) that unobserved local factors drive, at least in part, the cross-country variation in the dependent variable.[9] In most cases, researchers are painfully aware of potentially important variables that are missing from the analysis. Accounting for these missing variables is not atheoretical; it is simply careful science.

2.3 *Dynamics*

In general, there is little to guide researchers on what type of dynamic specification to employ when using TSCS. Hopefully, theory will be some guide, but theory is often not fine-tuned enough to point toward one particular specification. For the sake of reference later in the analysis, we list a few common specifications here. We limit our discussion to models that have up to only one lag in the dependent and independent variables and drop the unit index i. We consider the following six models:

$$\text{Static model}: \quad Y_t = \alpha + \beta_0 X_t + u_t, \tag{2}$$

$$\text{AR}(1)\text{ model}: \quad Y_t = \alpha + \beta_0 X_t + u_t; \quad u_t = \rho u_{t-1} + e_t, \tag{3}$$

$$\text{DL}(1)\text{ model}: \quad Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + u_t, \tag{4}$$

$$\text{LDV model}: \quad Y_t = \alpha + \beta_0 X_t + \gamma_1 Y_{t-1} + u_t, \tag{5}$$

$$\text{ARDL}(1,1)\text{ model}: \quad Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \gamma_1 Y_{t-1} + u_t, \tag{6}$$

$$\text{FD model}: \quad Y_t - Y_{t-1} = \beta_0 (X_t - X_{t-1}) + u_t \tag{7}$$

(AR: autoregressive; DL: distributed lag; ARDL: autoregressive, distributed lag[10]; FD: first difference).

[9] Although we have painted a stark picture of a trade-off between estimating sluggish variables and controlling for omitted variable bias with fixed effects, T. Plumper and V. E. Troeger (unpublished data) suggest a way that we may be able to get the best of both worlds. They propose a three-stage estimator that helps to control for unit heterogeneity and still estimate sluggish variables. They present a Monte Carlo work, which shows that under certain conditions, their proposed estimator outperforms traditional estimation with fixed effects or random effects. Their preliminary results for their proposed estimator are encouraging and suggest that the estimator may provide a better option to dealing with the traditional variance versus bias trade-off.

[10] The ARDL(1,1) model includes other models as special cases. These include several AR models including the Koyck model, the adaptive expectations model, and the partial adjustment model. Also, the AR(1) model shown above can (after a little algebra) be written as a special case of an ARDL(1,1) model.

There may be valid theoretical reasons for picking one dynamic specification over another. Lacking such a justification, however, there is no logical reason why the LDV model championed by B&K should be considered more plausible than any of the other dynamic models.[11] Just because one expects the value of Y_{it} to be near the value of Y_{it-1} is not a sufficient argument for selecting the LDV over the other dynamic models. *All* the models discussed above share a common feature: they capture the contemporaneous effect of X_t on Y_t, $(\partial Y_t/\partial X_t)$, through the parameter β_0, though the long-run dynamics will differ. Without further theoretical justification, each of these models seems as plausible as the next, but they differ in their levels of generality.[12]

2.4 *The Dynamic Panel Model*

In the preceding sections, we considered the issues of unit heterogeneity and dynamics separately, even though it is difficult to disentangle dynamic issues from heterogeneity in practice. A model that has both unobserved unit effects and an LDV is commonly referred to as a dynamic panel model (DPM):

$$Y_{it} = \lambda Y_{it-1} + \beta X_{it} + \alpha_i + u_{it}. \tag{8}$$

Before discussing estimation of the DPM, it is important to note that, intuitively, the LDV term and the unit effect work to anchor the overall level of the dynamic process in Y_{it} that is occurring. Some might even argue that the LDV model solves the heterogeneity problem (since a "high" unit effect is going to be associated with a high level of Y_{it-1}). However, this need not be the case. If, for simplicity, we assume that the X_{it} process is stationary and that $E(X_{it}) = \theta_i$, then Y_{it} converges in the long run to $(\alpha_i + \beta\theta_i)/(1 - \lambda)$. Unless the effect of the X variable dominates the unit effects (meaning α_i is very small relative to $\beta\theta_i$), significant unit heterogeneity will lead each series to converge to a different level. Indeed, a primary reason to estimate the DPM model, as opposed to the simple LDV model, is to capture this important variation in the long-run dynamics. The interpretation of unit effects in the DPM is different from the simple static model illustrated in Fig. 1, but the problem of unit heterogeneity still exists when the dynamics are captured by an LDV.

Unfortunately, the problem with the DPM is that OLS is no longer unbiased or consistent as long as T is finite (and dropping the LDV or the unit effects simply results in a different kind of bias).[13] Nickell (1981) has derived the exact formula for the bias.[14] As T grows, the bias can be shown to disappear, but this is little help for most TSCS studies, where T is seldom more than 30. Several asymptotic (in N) estimators have been proposed in the literature,[15] including a generalized method of moments-based estimator of Arellano and Bond (1991), a "corrected" least-squares dummy variable estimator of Kiviet (1995), a "nearly unbiased" estimator of Carree (2001), and a maximum likelihood estimator by Hsiao, Pesaran, and Tahmiscioglu (2002), but again this is of little help for TSCS studies where N is usually small.

Currently, work is underway to try to establish the behavior of these different estimators in TSCS data sets. The initial results are quite encouraging. Monte Carlo analysis by

[11] Furthermore, the LDV model is biased and inconsistent in the presence of autocorrelation. Sometimes, the LDV model can eliminate the autocorrelation, but, if it does not, then the coefficient estimates will be biased (this is somewhat ironic, in that autocorrelation by itself does not cause bias). B&K acknowledge this in their 1996 article, though we show that many researchers using the LDV actually do not test for autocorrelation.

[12] Of course, TSCS analysis should deal with all the issues that confront time-series analysis more generally, such as unit roots, cointegration, multiple lag structures, etc.

[13] This has been known at least since the Monte Carlo studies of Nerlove (1971).

[14] A more general discussion of the potential bias in DPMs can be found in Kennedy (1998, 149–50).

[15] The literature on this topic is large. Wawro (2002) gives an excellent review.

Judson and Owen (1999), C. Adolph, D. M. Butler, and S. E. Wilson (unpublished data), and N. Beck and J. N. Katz (unpublished data) suggests that the bias of FEM estimates of β is quite small in the types of data sets typically used in political science. However, the bias can be significant for estimates of λ, the coefficient of the LDV. In most studies in political science, however, λ is of little direct interest (except when estimating long-run dynamics is desired). Given that the bias associated with FEM estimation of β may be small, applying these asymptotic estimators may create more problems than they solve, and the simple FEM estimates of equation (8) are likely to be the best choice.

2.5 *Samples of Repeated Observations*

A fundamental question in using TSCS data is whether the repeated observations within a country can be considered as legitimate (see Kittel 1999). For example, consider investigating the effect of regime type on the provision of public goods with data on 20 countries. Suppose now that we obtain 20 years of data for these countries, even though regime type never changes within those 20 years. The *t* statistic on the regression equations are certainly going to rise, but is this data inflation really legitimate? Why not take monthly observations for each of these 20 countries, then we would have 4800 data points, and surely all our estimates would be statistically significant. So why, in turn, is the sample of 400 legitimate, rather than using between-country comparisons in the sample of 20? In short, if we have 20 countries in our data set, we have 20 countries, not 400. We are only justified in including multiple years of data if there is enough change within the variables and relationships we are examining to consider the individual years as distinct observations rather than the same observation copied over again.[16] This topic has received little attention in the literature.

3 Does Method Matter?

3.1 *A Methodological Review*

Of those papers that have cited B&K (1995) and/or B&K (1996), we identified 195 studies that present original analyses using linear panel data methods. In this section we summarize some of the key methodological features of this literature as they relate to our critiques. We restrict our review to studies that are published in political science journals indexed in the Social Science Citation Abstract as of July 1, 2005. We do not analyze nonlinear models (including probit or logit) nor do we consider the few studies that use instrumental variable estimation or other methods. In Table 1 we summarize our review of these studies along a number of criteria.[17] We are looking for two central features of TSCS data analysis: whether the authors consider unit heterogeneity and whether they consider dynamic specifications beyond the basic LDV or AR(1) models (including testing for autocorrelation).[18]

[16] We note here that a different type of problem can occur if the data are not measured frequently enough, namely that of temporal aggregation, which can lead to several types of incorrect inferences. If the true generating process is such that observations occur at frequent intervals, then the empirical model should include observations measured at the same frequency. Really, temporal aggregation is the same type of problem we are discussing here, but in reverse—namely, masking legitimate observations with temporal aggregates. Temporal aggregation has long been studied in the econometric literature but has been mostly neglected in political science. Important exceptions include Freeman (1989), Alt, King, and Signorino (2001), and Shellman (2004).

[17] The complete data for this review can be found in Appendix A, which is available at the *Political Analysis* Web site, as well as at http://fhss.byu.edu/Faculty/sew22/papers.

[18] For purposes of Table 1, we chose not to count the AR(1) model (seen in practice in the form of a Prais-Winston correction) as an alternative dynamic model since it focuses on correction of the error terms rather than the equation specification directly. We do report the number of articles using a Prais-Winston correction alongside those using an LDV (but not higher order lags).

Table 1 Summary of key methodological issues in published political science TSCS studies

195 Studies reviewed
Unit heterogeneity
77 (39.5%): number that consider unit heterogeneity
5 (6.5%): number which test and reject unit heterogeneity before excluding
72 (93.5%): number which report unit heterogeneity
118 (60.5%): number that do not consider unit heterogeneity
Dynamics
43 (22.1%): number that use models with no dynamics
3 (6.9%): justification provided for not using dynamics
40 (93.1%): no discussion provided
101 (51.8%): number that use the LDV and/or PW but not any alternative dynamics
75 (74.3%): number using an LDV (with or without PW correction)
Reasons given for including the LDV
20 (26.7%): theoretical
52 (69.3%): to correct for autocorrelation
45 (60.0%): recommended by B&K
22 (29.3%): number that test for autocorrelation
26 (25.7%): number using only PW correction
5 (5.0%): number using both an LDV and the PW correction
51 (26.1%): number that use or consider alternative dynamics
12 (23.5%): number that test for autocorrelation

Note. This is based upon articles found doing a search for articles citing B&K (1995 or 1996) on the Social Sciences Citation Index on July 1, 2005. This list only represents those articles found in that search which used linear models with TSCS data. Articles using other methods were not included in this table. PW = Prais-Winston.

Our analysis gives each paper a strong benefit of doubt. All we are asking at this point is whether authors even *consider* heterogeneity and dynamics. Our treatment of the word "consider" is also quite liberal, and we count very brief and tangential discussions of these issues as full consideration.

On the issue of unit heterogeneity, we found that only 39.5% of the studies reviewed considered unit effects. Encouragingly, 93.5% of those studies that do consider unit effects end up reporting the results from those regressions. Nonetheless, a majority of studies do not even consider (much less test) for the presence of unit effects. It may be that some of these studies test and reject unit effects, but, if this is the case, this useful information is not reported in the analysis.

On the issue of dynamics, we found that 22.1% of studies have models with *no* dynamics. Of those 43 studies, 40 (93.1%) provide no justification for why they are ignoring dynamic issues. We find that 51.8% of studies do not consider any model beyond the basic LDV or AR(1) models, with the vast majority of those articles reporting the LDV model encouraged by B&K (38.5% of the total). Of those who use the LDV model, only 29.3% test for autocorrelation, even though autocorrelation in the presence of LDVs causes biased estimation of the coefficients. We also note the reasons given for using the LDV model, with citations to B&K or as an autocorrelation correction constituting the most important reasons. Only 26.7% of studies cite theoretical reasons for choosing the LDV specification. Finally, only 26.1% of all the studies consider alternative dynamic models beyond the LDV or AR(1) models, and only 23.5% of those test for autocorrelation.

Careful studies using TSCS should consider unit heterogeneity and alternative dynamic specifications and test for autocorrelation. Our summary shows that of the 195 studies,

only 53 studies consider both unit heterogeneity and some kind of dynamic structure. However, only 25 of those use or consider any dynamics beyond the simple LDV model. Finally of this number, only seven also include tests for autocorrelation. Thus, less than 5% of studies cover the basic criteria that we have laid out. This does not mean, of course, that the other 95% of the studies are invalid. But it does imply that they are incomplete in some important way, especially since our criteria our minimal and do not include additional important specification issues such as endogeneity or higher order lag structures. Many studies do have thoughtful analysis of methods, but in general, we find a lack of attention to specification issues and a failure to adequately consider well-known models found in the literature.

The introduction of PCSEs was a helpful advance, but we suspect that the problems researchers tend to ignore are far more serious than the problems corrected with the PCSEs. In many cases, PCSEs lead to the same inference as the OLS standard errors, which probably entices some researchers to believe that their results are robust. We conclude from our extensive reading of the literature that more than a few researchers are using B&K (1995) as a complete and authoritative guide to conducting TSCS analysis. Far too much of the research neglects the long-existing literature on panel data methods, and almost none of it acknowledges the potential bias associated with panel data models as we discussed in Section 2.4.

3.2 *Robustness*

We next examine whether published results are sensitive to alternative specifications. To do this, we chose eight published studies and reanalyzed the data incorporating unit effects and alternative dynamic specifications. We did not choose studies for analysis randomly nor do we make broad claims about their representativeness. Of the studies we reviewed earlier, we picked 20 from the top journals in political science, of which we were then able to obtain the data for eight studies for replication. We were somewhat biased toward those studies that had data available online, but in most cases, we picked pieces that we thought were of high quality, those pieces recommended by colleagues, and those that, more or less, followed the B&K method. By and large, the authors were helpful in providing their data. Before proceeding, it is important to note that these studies are not the worst offenders of the problems that we have discussed. For example, one study tests an alternative dynamic specification,[19] two studies test the effect when the LDV is dropped from the model,[20] three studies give theoretical reasons for including the LDV,[21] and five studies test for serial correlation.[22] Still, seven of the articles did not consider alternative dynamic models, and none of them test for unit effects.[23]

A typical article reports a few different models that contain many different variables. We wanted to highlight the effects of our tests on the central conclusions of the paper. For each of the eight articles we replicated, the challenge was to pick results that were both representative and relevant to our critiques. We chose one or two specifications that both capture a central point of the authors' analysis and that are simple to interpret. Thus, we

[19]Zahariadis (2001) included lagged independent variables along with the LDV, giving theoretical reasons for both.
[20]Cox, Rosenbluth, and Thies (1998) and Pickering (2002).
[21]Reich (1999), Moene and Wallerstein (2001), and Zahariadis (2001).
[22]Poe and Tate (1994), Cox, Rosenbluth, and Thies (1998), Moene and Wallerstein (2001), Zahariadis (2001), and Pickering (2002).
[23]Although none of the authors test for fixed effects, both Reich (1999) and Moene and Wallerstein (2001) do mention them briefly.

Table 2 Published findings replicated and analyzed

Article	*Page no.*	*Table column*	*Dependent variable*	*Important independent variables*
Cox, Rosenbluth, and Thies (1998)	466	1.1	Total expenditures per elector	Victory margin
Cox, Rosenbluth, and Thies (1998)	467	2.2	Voter turnout	Total expenditure, victory margin, percent men, percent urban, percentage of population under 15
Hood, Kidd, and Morris (2001)	611	1.1	Unadjusted LCCR scores	GOP strength, black electoral strength
Moene and Wallerstein (2001)	869	1.2	Government spending on income insurance	Inequality (90/10)
Moene and Wallerstein (2001)	869	1.5	Government spending on income insurance	Inequality (90/50), inequality (50/10)
Pickering (2002)	328	2.2	Foreign military intervention scale	War experience, war experience squared
Poe and Tate (1994)	861	1.4	Personal integrity abuses	Democracy (Van Hanen)
Reich (1999)	743	1.3	Seniorage	First democratic government
Reich (1999)	743	1.4	Seniorage	Democracy for less than 10 years
Saideman et al. (2002)	119	1.1	Protest	Regime type, first election, federal system, proportional democracy, enduring regime, young democracy
Zahariadis (2001)	613	2.1	Total aid	Research and development, research and development squared, job gain

Note. LCCR, Leadership Conference on Civil Rights.

did not choose models with interaction terms, though these models were theoretically interesting. In this section we have the space to only briefly summarize the results. The most important restriction is that we concentrated on only those coefficient estimates that we determined, a priori, were central to the author's arguments. Table 2 lists the papers, models, and relationships analyzed. Appendix B contains the coefficient estimates of the variables that are the focus of our analysis. Appendix C contains the complete regression results.[24]

As previously noted, the robustness analysis we conduct here is narrow in scope. Much more ambitious sensitivity analysis could be conducted both in terms of alternative specifications and alternative variables that might be included. Leamer (1983, 1985) introduced the idea of "extreme bound analysis," which tests for the inclusion of other variables present in the literature on the estimates of the coefficients of interest.[25] This requires, of course, that other potential variables are available in the literature. Our approach here is in the spirit of "general-to-specific" modeling, sometimes referred to as the Hendry approach or the London School of Economics approach.[26] Though we are not specifying very general models and then reducing them, we are implicitly assuming that the alternative specifications we

[24]Appendices B and C are found online at the *Political Analysis* Web site.
[25]An important application of extreme bound analysis to TSCS data is a paper by Levine and Renelt (1992).
[26]A large literature exists on this type of specification testing. See, for example, Hendry (1995) and Campos, Ericsson, and Hendry (2005).

test are part of a more general model. But since other alternative specifications exist other than the ones we test, we do not have sufficient evidence to say which is the best model.

3.2.1 Unobserved heterogeneity

Table 3 reports the results of our replications and reanalysis when the selected models are estimated with and without unit effects. Since we are interested in what happens to the sign, magnitude, and statistical significance of coefficient estimates in these studies, we summarize the findings in terms of whether the findings are strengthened, unaffected, weakened, or reversed, which can be interpreted as corresponding to the four different scenarios in Fig. 1. We refer to a finding as strengthened or weakened if the FEM results in a change in magnitude of at least half of a standard error, as measured by the PCSE of the original results. In all cases, those findings that are statistically significant are noted in italics on the table. Further analysis of changes in magnitude and significance are found in Appendix B.

The dominant story of Table 3 is the general instability of regression coefficients as unit effects are added to the model. Although some findings are left unaffected, considerably more are altered substantially by the FEM. It is true that the consequences of heterogeneity are relatively benign in some cases. For example, in the Hood, Kidd, and Morris (2001) analysis of Civil Rights voting by Southern Senators, the basic findings that both the electoral strength of the Black electorate and the Republican Party (GOP) has pushed Democrats to the left are confirmed, but the relative importance of the GOP is increased by a factor of 3, whereas the effect of Black electoral strength falls slightly. We also conclude that Reich's (1999) analysis of the effect of democratic transition on seniorage (1999) and Poe and Tate's (1994) analysis of democracy and governmental repression are essentially robust to the inclusion of unit effects, though there are some differences in coefficient magnitude and t statistic that are noteworthy.

The other papers show much more extreme consequences of unit heterogeneity. For instance, Pickering (2002) reports a U-shaped effect of past conflict on military intervention scale (the number of troops deployed), implying that as the number of successive wins or successive failures increases, the number of troops deployed increases. The FEM analysis shows just the opposite—the effect of wartime experience is now an *inverted* U, with a maximum at 0.80 prior wars. This implies that each additional win or loss in a streak (after the initial one) decreases the scale of the next military intervention. Similarly, Moene and Wallerstein (2001) estimates of the effect of inequality on government spending for insurance against loss of income are reversed under the FEM model, and in one case the FEM coefficient is even statistically significant.[27]

Though space does not permit a substantive critique of each paper, we note that all the remaining papers exhibit a notable nonrobustness in key findings. The Saidemen et al. (2002) results show even a stronger effect of regime type than that the authors find, but find opposite effects for the duration of regime (the FEM finds that enduring regimes have less protest, not more). Most of the findings from Zahariadis (2001) are weakened and the Cox, Rosenbluth, and Thies (1998) results are a mixed bag of strengthening and weakening of the key coefficients.[28]

[27] Interestingly, the authors claim that their results are "destroyed" if fixed effects are included, but they claim that there are not sufficient data to test the FEM. We agree, but should not this mean that the data are also insufficient to accept their estimates with the fixed effects deleted, especially since the two models have entirely different conclusions?

[28] Neither of these two papers actually reported PSCEs in their published results (Cox et al. give their reasoning in footnote 49), but we use the PSCEs to preserve consistency. Some of the Cox, Thies, and Rosenbluth results that are statistically significant with OLS standard errors are not significant with PCSEs.

Table 3 Robustness of results to inclusion of unit effects

Article	*Dependent variable*	*Independent variables*
Estimates that increase in magnitude		
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Percent men
Hood, Kidd, and Morris (2001)	Unadjusted LCCR scores	*GOP strength*
Saideman et al. (2002)	Protest	*Regime type*
Estimates that remain unchanged		
Cox, Rosenbluth, and Thies (1998)	Total expenditures per elector	*Victory margin*
Cox, Rosenbluth, and Thies (1998)	Voter turnout	*Victory margin*
Reich (1999)	Seniorage	*First democratic government*, democracy for $<$10 years
Saideman et al. (2002)	Protest	First election
Estimates that fall in magnitude		
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Total expenditure
Hood, Kidd, and Morris (2001)	Unadjusted LCCR scores	*Black electoral strength*
Pickering (2002)	Foreign military intervention scale	War experience
Poe and Tate (1994)	Personal integrity abuses	Democracy (Van Hanen)
Saideman et al. (2002)	Protest	Federal system, proportional democracy
Zahariadis (2001)	Total aid	R&D, R&D squared
Estimates where sign is reversed		
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Percent urban, percent population $<$15
Moene and Wallerstein (2001)	Government spending for income insurance	Inequality (90/10), *inequality (90/50)*, inequality (50/10)
Pickering (2002)	Foreign military intervention scale	War experience squared
Saideman et al. (2002)	Protest	Young democracy, *enduring regime*
Zahariadis (2001)	Total aid	Job gain

Note. These results represent the effect of adding fixed effects to the original model. The independent variables that are statistically significant at the 0.05 level after the inclusion of FE are italicized. For all the studies we used PCSEs to determine statistical significance, including Cox, Rosenbluth, and Thies (1998) and Zahariadis (2001) which did not use PCSEs in their initial studies. Estimates that increase (fall) in magnitude refer to those variables where the coefficient magnitude increased (decreased) by at least half the initial standard error. If the estimate did not change by more than half the initial standard error, we classified it as unchanged. The final category includes variables where the sign of the coefficient changed. Appendix A includes the regression results for the major independent variables of each study. Appendix B, available online, contains regression results for all the variables in these regressions.

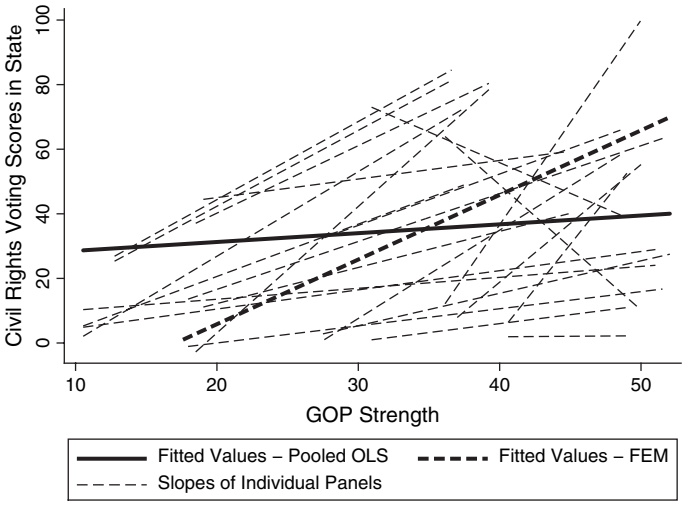

Fig. 2 Civil rights voting among southern senators, 1970–1996. Data provided by Hood, Kidd, and Morris (2001).

The estimates summarized in Table 3 include an LDV, since the original model included one as well. In the Appendix B, we also note the impact of unit effects for the static model (no LDV). In comparing the consequences of fixed effects for both the static and the LDV model, two distinct phenomena are apparent. First, the static model coefficients are almost always larger in magnitude than the LDV coefficients, as we would expect. Since the LDV model uses the dependent variable to explain itself, it is hardly surprising that the effects of other variables are reduced. However, the second result is that the introduction of unit effects has nearly the same effect on the coefficient estimates regardless of whether one starts with the static or the LDV model. In other words, the LDV may be more a more conservative approach than the simple static model, but the consequences of unobserved heterogeneity are found in both the static and the LDV framework. In other words, the LDV approach of B&K is not a solution to the unit heterogeneity problem, at least among the coefficients we have examined.

To conclude our analysis of unit heterogeneity, we provide two visual illustrations of how coefficient estimates can be sensitive to the inclusion of fixed effects. Figure 2 shows a relationship analyzed by Hood, Kidd, and Morris (2001). They find in their analysis that the GOP strength pushed Democratic Senators to the left in terms of their civil rights voting record. As shown in Appendix B, including fixed effects increases the magnitude of this effect by 272%. In Fig. 2, the bold solid line represents the pooled OLS estimate, and the bold dashed line represents the FEM estimate. The regular dashed lines represent the simple regressions with respect to GOP strength for each of the 22 Senate seats used in their analysis. It is not hard to see why the FEM model shows stronger results: most of the 22 units exhibit a strong positive relationship between the two variables. Pooled OLS imposes a common intercept that, in this case, masks the positive slope clearly present in the data. Allowing the slopes to vary with the FEM model allows the positive relationship to be revealed (though we caution that this general upward trend might be due to unobserved factors that trend upward over time).

Figure 3 tells an entirely different kind of story—one in which pooling suggests a relationship that might not be there under closer examination. One of the pooled OLS

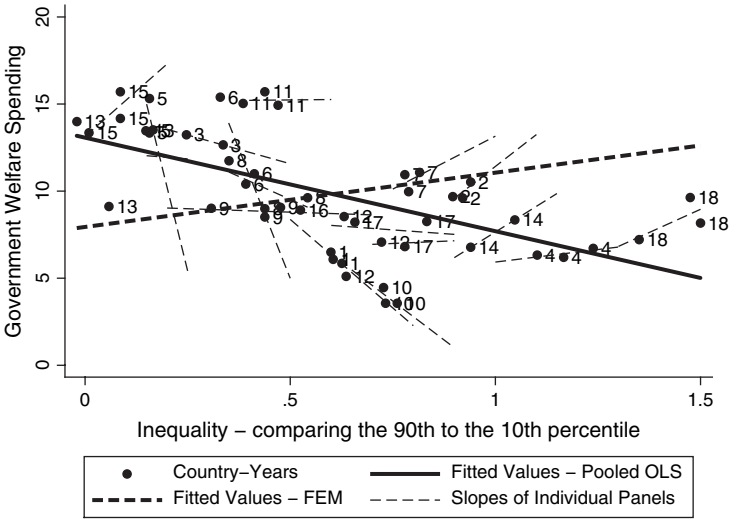

Fig. 3 Inequality and welfare spending in OECD. Data provided by Moene and Wallerstein (2001).

estimates of Moene and Wallerstein (Table 1, column 2) indicates a negative relationship between income inequality and social welfare spending in Organization for Economic Cooperation and Development (OECD) countries. Looking at the scatterplot of points in Fig. 3, this conclusion seems reasonable, given the general negative trend seen in the data. However, our FEM estimation of the same model (see Appendix B) finds a positive effect, though it is not statistically significant.

Given the small sample size, we do not place much stock in the FEM results shown in Fig. 3 nor would we claim that the FEM results are correct and the pooled OLS results are incorrect. But using the FEM as a diagnostic tool reveals that the authors' results are driven primarily by the cross-sectional variation present in their data. In contrast, the thin dashed lines of Fig. 3 show that—*within countries*—only a few of the simple relationships are actually negative. Would not the authors' theory imply a negative relationship within countries as well as across countries? The problem with the pooled OLS results (including the reported standard errors) is that they assume 50 *independent* data points when, in fact, there are only 18 countries. Thus, their reported results overstate the true cross-sectional relationship and ignore the inconsistent longitudinal patterns within the data.

3.2.2 Alternative dynamics

Above we identified the consequences of adding unit effects to the basic LDV model of B&K. In this section, we assume that unit effects are not relevant and explore the consequences of estimating models with alternative dynamic structures. In short, for each case we compare the estimate of five alternative dynamic models discussed in Section 2.3 with the LDV model of B&K. We argued earlier that each of these models is a plausible alternative to the LDV model, though theoretical reasons may rule out particular dynamic specifications in some cases. However, we note two important caveats. First, because a finite sample will always generate nonzero correlations between the explanatory variables and nonzero regression coefficients, there will always be some variance in the estimates of β_0 for a given sample. These incidental effects may be quite large in small samples. Second, the six models we estimate are far from comprehensive. Higher order

lag structures can (and should) be explored, for instance. But if the published results we examine are not stable across the simple linear models we propose, they are unlikely to be robust to other alternative specifications and estimation approaches.

Table 4 summarizes the results of this analysis as follows. First, we report the range of coefficient estimates represented in terms of actual values and in terms of the number of standard errors (using the standard error from the PCSE from the LDV model). For example, the effect of R&D on total aid in the Zahariadis (2001) analysis gives estimates ranging from -2.317 to -0.517. Since the PCSE is 0.423, this range of estimates is equivalent to 4.3 standard errors.[29] The variation is also expressed in terms of the percentage of deviation from the mean estimate. The minimum and maximum deviations as well as the median are reported.

We have organized the results into categories based on the published results and the results of our analysis. The first three sets of coefficients are reported by the authors to be statistically significant. First we report findings that are "robust," which means that they all have the same sign, a relatively low range of coefficients, and a high number (five or six) of estimates that are statistically significant. We find five estimates from three studies that satisfy these criteria. An example of a particularly robust finding is the effect of victory margin on voter turnout (Cox, Rosenbluth, and Thies 1998), where a small range of statistically significant coefficient estimates ranges between -0.104 and -0.082. The second category consists of "weakly robust" findings, which have at least three significant findings with no sign reversals (except in one instance).[30] The third category of findings includes those that are "nonrobust." In this case, variations in sign are common, the range of estimates is high, and statistical significance is relatively uncommon. In many cases, the methods even obtain significant results of different sign.

The next two categories concern findings reported in the published studies as insignificant. A "robust nonfinding" occurs if all the other methods yield a small range of insignificant effects close to zero. A "weakly robust nonfinding" is insignificant, but the range of estimates is so large that it is hard to be confident that the effect is actually zero. This is particularly likely to occur in the case of small samples such as Moene and Wallerstein (2001). It is possible to have a "nonrobust nonfinding" as well, which would occur if alternative specifications led to strong, contradictory results, but none of the estimates fell into this category.

This analysis reveals, as with the FEM results earlier, that simple methodological alternatives can have profound results. Although many of the estimates were either robust or weakly robust, six of the eight studies had findings that were not robust, and all eight had at least one finding that was only weakly robust. Furthermore, this analysis does not include the multiple other variables in the regression models (see Appendix C for complete results). Were we to extend this analysis one step further by turning the six dynamic models into 12 by including unit effects within each variation, the variance in estimates and the resulting uncertainty regarding some of the published findings would clearly become even larger. And, finally, even though many results are classified as either robust or weakly robust, the range of estimates is in most cases quite high, usually far outside the 95% confidence intervals that are associated with the LDV estimates. To the extent that we care about what

[29] $[(2.317–0.517)/0.423] \approx 4.3$.

[30] The exception to this is the effect of campaign expenditures on voter turnout in the Cox, Rosenbluth, and Thies model. In this case, it is only the simple static model that gives a negative coefficient, whereas the other models find positive and generally significant effects.

Table 4 Robustness of results to alternative dynamic specifications

Article	Dependent variable	Independent variable	Range of estimates: Coefficients	Range of estimates: No. of standard. errors	Percentage of deviations from mean estimate: Minimum	Percentage of deviations from mean estimate: Maximum	Percentage of deviations from mean estimate: Median	No. of times variable was significant
Coefficients reported as statistically significant using the LDV model								
Robust findings								
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Victory margin	[−0.104, −0.082]	1.2	0	16	5	6
Saideman et al. (2002)	Protest	Federal system	[0.123, 0.322]	4.2	23	47	53	5
Saideman et al. (2002)	Protest	Proportional democracy	[−0.818, −0.128]	14.6	17	77	38	6
Saideman et al. (2002)	Protest	Regime type	[0.019, 0.071]	8.3	14	60	24	6
Weakly robust findings								
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Expenditures	[−0.062, 0.065]	6.5	24	300	70	4
Hood, Kidd, and Morris (2001)	Unadjusted LCCR scores	Black electoral strength	[0.080, 2.631]	4.1	17	91	46	3
Pickering (2002)	Military intervention scale	War experience	[0.069, 0.164]	3.4	31	41	32	3
Poe and Tate (1994)	Personal integrity abuses	Democracy	[−0.026, −0.003]	7.6	17	114	64	3
Reich (1999)	Seniorage	Democratic <10 years	[0.871, 3.269]	4.3	13	72	24	3
Zahariadis (2001)	Total aid	R&D	[−2.317, −0.517]	4.3	11	74	66	4
Nonrobust findings								
Cox, Rosenbluth, and Thies (1998)	Total expenditures	Victory margin	[−0.222, 0.017]	5.9	58	197	91	5
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Percent population <15	[0.419, 1.589]	20.9	2	216	62	2
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Percent men	[−0.721, 5.445]	11.3	40	259	140	2
Cox, Rosenbluth, and Thies (1998)	Voter turnout	Percent urban	[−0.764, 0.065]	49.4	25	410	90	3
Hood, Kidd, and Morris (2001)	Unadjusted LCCR scores	GOP	[−1.391, 0.793]	9.4	316	453	412	3
Moene and Wallerstein (2001)	Income insurance	Inequality (50/10)	[−4.696, 2.544]	18.1	87	580	197	2
Moene and Wallerstein (2001)	Income insurance	Inequality (90/10)	[−4.442, 3.222]	24.2	197	597	247	3

Continued

Table 4 (continued)

Article	Dependent variable	Independent variable	Range of estimates: Coefficients	Range of estimates: No. of standard. errors	Percentage of deviations from mean estimate: Minimum	Percentage of deviations from mean estimate: Maximum	Percentage of deviations from mean estimate: Median	No. of times variable was significant
Pickering (2002)	Military intervention scale	War experience squared	[−0.170, 0.092]	17.9	116	867	153	3
Reich (1999)	Seniorage	First democratic government	[0.724, 1.910]	2.6	11	60	26	2
Zahariadis (2001)	Total aid	R&D squared	[−0.021, 0.448]	3.8	24	109	130	3
Coefficients that were reported as statistically insignificant using LDV model								
Robust nonfindings								
Saideman et al. (2002)	Protest	First election	[−0.198, 0.094]	2.5	78	323	36	0
Zahariadis (2001)	Total aid	Job gain	[−0.017, 0.005]	2.9	48	199	20	0
Weakly robust nonfindings								
Moene and Wallerstein (2001)	Income insurance	Inequality (90/50)	[−23.74, −0.675]	24.2	24	323	80	2
Saideman et al. (2002)	Protest	Enduring regime	[−0.192, 0.105]	5.8	17	437	147	0
Saideman et al. (2002)	Protest	Young democracy	[−0.028, 0.101]	2.1	16	172	196	0

Note. The B&K method uses pooled OLS with an LDV and PCSEs. It should be noted that in their initial studies, Cox, Rosenbluth, and Thies (1998) and Zahariadis (2001) did not use PCSEs and Saideman et al. (2002) also used the Prais-Winston (PW) in addition to the basic B&K method; however, to be consistent, we used the B&K method when performing replications for this table. We tested the robustness of the B&K method results by running five other model specifications: pooled OLS without the LDV, PW without the LDV, lagged independent variables (LIVs), LIVs and LDV, and first differences. We used PCSEs in all the models. For each of the estimates, we list the range (minimum and maximum) of coefficient values for the specifications we tested and the number of standard deviations, using the standard error from the B&K method, that the range covers. The variation is also expressed in terms of the percentage of deviation from the mean estimate, with the minimum, maximum, and median deviations reported. The final column lists how many of those specifications gave statistically significant results at the 0.05 level. For all the studies we used PCSEs to determine statistical significance. The first three categories are those that were statistically significant using the B&K method. A given result was considered a robust finding if all the estimates had the same sign, there was a relatively low range of estimates and a high number of estimates that were statistically significant. To be considered a weakly robust finding, the estimates had to be statistically significant in at least three of the specifications and had to keep the same sign. The nonrobust findings included variables where sign reversals were common, statistical significance was uncommon, and variation in range of coefficients was high. The final two categories were those that were reported as statistically insignificant in the published reports. A robust nonfinding indicates that all the other specifications yielded a small range of insignificant effects close to zero. A weakly robust nonfinding is a result that was reported as insignificant, but the range of estimates is so large that we cannot be confident that the effect is actually zero.

the coefficient estimates actually are (rather than if they are merely different from zero), almost all the estimated effects reported in the literature give us cause for concern.

3.3 *Lessons Learned*

So what can be learned from the exercise above? The first lesson is good news: a few of the studies contained results that were robust to both controls for heterogeneity and to alternative dynamics. In some cases, adding unit effects has little impact on the coefficient estimates; in others, it actually strengthens them. But the larger lesson is less optimistic. We find that many of the conclusions reached in the published studies we examined are highly contingent on the method used to obtain them. The nonrobustness in some cases is modest, such as reduction in the magnitude of a coefficient or the failure to obtain statistical significance in all the alternative specifications. In other cases the nonrobustness is stark, with different specifications leading to opposite and statistically significant results and a high variance in estimates across methods. Our assessment is that, in general, the findings from the TSCS studies we examined can only be regarded as frail. Of course the lesson that method matters is an old one, but it is particularly appropriate to emphasize in the case of TSCS data analysis.

We need to stress here that we are not arguing for any one specification over another nor are we championing one estimator over another. Findings that we designate as nonrobust may prove to be correct, and findings that appear solid may, in fact, be all wrong. It is also true that given the known bias of dynamic panel data models discussed in Section 2.4, all the estimates we obtain are potentially biased, since they all include both unit effects and LDVs. This theoretical result, in itself, should make analysts wary of many of the published works in this area. Combining the theoretical potential for bias with both the sensitivity analysis we conducted above and the general lack of attention to specification issues in the published literature (as shown in Section 3.1) gives ample justification for the claim that published studies using the B&K method deserve further scrutiny.

4 Moving Forward

In the exercise above, our desire is not to denigrate the research that has been undertaken in the past. Much of it is very careful and insightful. But we hope to have shown the need for extensive sensitivity testing as part of the research process. It may be that including unit effects or allowing for alternative dynamic specifications other than the simple LDV model will significantly challenge central findings. Given a field in which everyone is painfully aware that theoretical concepts sometimes have weak empirical analogues and where data collection is often error-ridden, highly aggregated, or otherwise problematic, *the bar for confirming theories with regression analysis should be very high.*

Given the challenges involved in estimating dynamic panel data models, especially with small data sets, we do not think that a set of "best practices" has been developed. Certainly, we agree with Kittel (1999) that the pooled OLS analysis for many of the data sets available in political science is "less impressive than its advocates suggest" (p. 245). But given that researchers are undoubtedly going to keep using variations of the B&K method in the future, a few limited recommendations are in order. This advice is hardly pathbreaking, but our analysis suggests that many researchers are estimating TSCS models with relatively little attention to the significant challenges such analysis raises.

The most fundamental question a TSCS researcher can ask is whether repeated observations on the same unit of analysis really constitute legitimate observations. We discussed earlier how the distinction between cases and observations affects the validity of inference

with a given sample. In other words, at what frequency (yearly/quarterly/monthly?) is it appropriate to make repeated observations of the same analytical unit and still consider those observations as legitimate? The answer clearly depends on whether the within-country variation in the dependent variable is "sufficient" and whether that variation can be explained by variation in the independent variables. Since time-invariant and slowly changing variables have, by definition, low variance, any inferences about them in pooled data sets are highly suspect. Indeed, if the frequency of observation is high enough, any unchanging variable can appear to be statistically significant.

If the answer to the above question is affirmative, then the next question is whether the observations can be treated as independent. In general, the answer in cross-country regressions will be no, rendering the pooled OLS technique of B&K invalid in most cases. A large literature exists for estimating dependent data. We have highlighted FEM, the simplest approach, but we stress as strongly as possible that we are not advocating a blind application of the FEM model (just as we eschew a blind application of the B&K approach). The FEM sweeps away the cross-sectional variation and leaves only longitudinal variation within countries. Because the cross-sectional variation may be very important and because sluggish variables may be of interest, the FEM will not be satisfactory in some cases. *But the inadequacy of the FEM in a particular instance does not mean the researcher should ignore unit heterogeneity and use pooled OLS*. As noted earlier, a new approach (T. Plumper and V. E. Troeger, unpublished data) based on decomposition of the FEM into time-invariant and residual components shows some promise as an alternative to the FEM. It is still the case, however, that unobserved variables that are responsible for unit heterogeneity cannot be easily disentangled from observable sluggish variables. In some cases, researchers must cope with the fact that a small TSCS data set with significant unit heterogeneity is of limited usefulness. With TSCS data, significant diagnostic evaluation (including, but not limited to, the type we performed in Fig. 2 and 3) is always in order, and the FEM is an important part of the diagnostic kit. Conducting and reporting tests of the pooling assumption and estimates from models with and without fixed effects should be a standard part of the diagnostic repertoire. Other common regression diagnostics, such as DFITS (as well as the FEM coefficients themselves), can be used within the FEM framework to identify influential countries, sets of countries, or groups of years within countries.

When we leave the confines of the simple static model and enter into the dynamic world, a Pandora's box of alternative models and approaches presents itself. A natural starting point is the ARDL(1,1) model, which has lags of both the dependent and independent variables in the model, though we do not want to diminish the importance of testing for higher order lags. Because of the numerous dangers involved in including an LDV, the researcher will be fortunate if the LDV can be excluded in favor of the simple DL(1) model (or better yet, the static model). In all cases, tests for autocorrelation should be conducted and reported in all dynamic models (and the static model as well). As reported earlier, LDVs will cause the FEM to be biased, but the bias seems to be relatively small on the X variables (which are the variables of interest), though significant bias can exist on the LDV coefficient. Furthermore, omitting relevant fixed effects from the dynamic models will likely cause even more serious bias.

We have left many important issues and methods largely untouched, including random coefficient models, cointegration, unit root testing, endogeneity and instrumental variables estimation, Bayesian hierarchical models, and spatial regression models, to name a few. An important part of the "a lot more to do" is exploring these additional issues. Large literatures exist in statistics and econometrics on these topics, but they are infrequently

mentioned in the applied political science literature that uses TSCS data. An unfortunate consequence of the rapid adoption of the B&K approach is that it induced a sizable group of researchers to neglect this large body of methodological literature. Although we certainly sympathize with applied researchers looking for a simple, robust, and widely accepted approach to estimating dynamic models, simple methods are not always appropriate for complex problems. The suggestions we make here are merely starting points toward more careful TSCS analysis.

5 Conclusions

Any readers who are still asking "Where is the fix?" have missed the point entirely. We have suggested several ways to improve practice in this area of research, but our central message is that dispensing simple prescriptions for complex problems can have unfortunate consequences. It would be convenient if the "simpler is better" mantra were actually true in the case of TSCS data. We endorse as much simplicity as possible, but this does not mean the B&K method is the *best* simple approach. The profession needs to come to grips with the fact that regression results with TSCS data can be exceedingly frail, that more than the usual amount of caution should be exercised, and that (difficult as it is for some to swallow) many data sets simply have too many limitations to use in a reliable fashion—regardless of the estimation method employed.

As we noted at the outset, the tale we have told here is more than just a critique of a particular approach or an analysis of particular type of data structure. It is also a critique of methodological advice giving and those who follow it. The simple B&K prescriptions—given with no discussion of major issues such as unit heterogeneity that have been present in the voluminous literature on dynamic panel data modeling that existed long before 1995—led numerous researchers to believe (or to at least act as if they believe) that the well-worn tool of pooled OLS with the "new" PSCEs tacked on constituted the state-of-the-art method ("what to do and not to do") for TSCS data. It is more than a little ironic that even though B&K's analysis focused on the danger of using estimators without fully understanding their properties, so many in the profession applied the B&K method without paying any attention to the simple textbook issues that we laid out in Section 2.

The recommendations we give above focus mostly on statistical analysis. But many of the statistical issues would be more straightforward if researchers had stronger theories to draw on when specifying their statistical models. It is also true that some of the theories that are tested at the cross-national level could be tested at the individual level, where data are more numerous. For instance, many studies (such as the Moene and Wallerstein piece we replicate here) explore preferences for state social insurance using nationwide social insurance levels as the dependent variable. Iversen and Soskice (2001), on the other hand, were able to construct empirical tests for their theory about social spending preferences at the individual level.

Of course microlevel tests often would not be available, and given the nature of TSCS data, statistical robustness will remain unattainable in many cases (20 countries = 20 countries!). Although continued statistical analysis and the development of better methods should proceed, researchers must be prepared for the answer that regression analysis simply will not provide reliable conclusions in some instances—a humbling fact relevant to regression analysis generally, we might add, not just in the TSCS context. This fact also points to the unavoidable conclusion that research in comparative politics and international relations must remain *qualitatively* rich. The rift between qualitative and quantitative analysts is, especially in the case of small panel data sets, counterproductive.

Political science seems particularly well poised, we think, to pursue a methodological agenda of marrying quantitative and qualitative methods (both enlightened by stronger theories), since neither mode of analysis on its own will be sufficient in many cases. How to make such a marriage work is some methodological advice from which we could all benefit.

References

Alt, James, Gary King, and Curtis S. Signorino. 2001. Aggregation among binary, count and duration models: Estimating the same quantities from different levels of data. *Political Analysis* 9:21–44.

Andrews, Donald W. K. 1986. A note on the unbiasedness of feasible GLS, quasi-maximum likelihood, robust and spectral estimators of the linear model. *Econometrica* 54:687–98.

Arellano, Manuel. 2003. *Panel data econometrics.* Oxford: Oxford University Press.

Arellano, Manuel, and Stephen Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies* 58:277–97.

Baltagi, Badi H. 2002. *Econometric analysis of panel data.* New York: John Wiley.

Beck, Nathanial. 2001. Time-series cross-section data: What have we learned in the past few years? *Annual Review of Political Science* 4:271–93.

Beck, Nathaniel, and Jonathan N. Katz. 1995. What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89:634–47.

———. 1996. Nuisance v. substance: Specifying and estimating time-series cross-section models. *Political Analysis* 6:1–36.

———. 2004. Random coefficient models for time-series cross-section data. Social Science Working paper 1205, California Institute of Technology.

Campos, Julia, Neil R. Ericsson, and David F. Hendry. 2005. *General to specific modeling.* Edward Elgar.

Carree, Martin A. 2001. Nearly unbiased estimation in dynamic panel data models. Working Paper.

Cox, Gary W., Frances M. Rosenbluth, and Michael F. Thies. 1998. Mobilization, social networks, and turnout: Evidence from Japan. *World Politics* 50:447–74.

Freeman, John R. 1989. Systematic sampling, temporal aggregation, and the study of political relationships. *Political Analysis* 1:61–98.

Hausman, Jerry A. 1978. Specification tests in econometrics. *Econometrica* 46:1251–71.

Hendry, David F. 1995. *Dynamic econometrics.* Oxford: Oxford University Press.

Hood, M. V., Quentinn Kidd, and Irwin L. Morris. 2001. The key issue: Constituency effects and southern senators' roll-call voting on civil rights. *Legislative Studies Quarterly* 26:599–621.

Hsiao, Cheng M. 1986. *Analysis of panel data.* Cambridge: Cambridge University Press.

Hsiao, Cheng, M. Hashem Pesaran, and A. Kamil Tahmiscioglu. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109:107–50.

Iversen, Torben, and David Soskice. 2001. An asset theory of social policy preferences. *American Political Science Review* 95:875–93.

Judson, Ruth A., and Ann L. Owen. 1999. Estimating dynamic panel data models: A guide for macroeconomists. *Economic Letters* 65:9–15.

Kennedy, Peter. 1998. *A guide to econometrics.* 4th ed. Cambridge, MA: MIT Press.

Kittel, Bernhard. 1999. Sense and sensitivity in pooled analysis of political data. *European Journal of Political Research* 35:225–53.

Kiviet, Jan F. 1995. On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics* 68:53–78.

Leamer, Edward E. 1983. Let's take the con out of econometrics. *American Economic Review* 73:31–43.

———. 1985. Sensitivity analysis would help. *American Economic Review* 75:308–13.

Levine, Ross, and David Renelt. 1992. A sensitivity analysis of cross-country growth regressions. *American Economic Review* 82:942–63.

Moene, Karl Ove, and Michael Wallerstein. 2001. Inequality, social insurance, and redistribution. *American Political Science Review* 95:859–74.

Nerlove, Marc. 1971. Further evidence on the estimation of dynamic economic relations from a time series of cross sections. *Econometrica* 39:359–82.

Nickell, Stephen. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49:1417–26.

Parks, Richard. 1967. Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated. *Journal of the American Statistical Association* 62:500–9.

Pickering, Jeffrey. 2002. War-weariness and cumulative effects: Victors, vanquished, and subsequent interstate intervention. *Journal of Peace Research* 39:313–37.

Poe, Steven C., and C. Neal Tate. 1994. Repression of the human right to personal integrity in the 1980s: A global analysis. *American Political Science Review* 88:853–72.

Reich, Gary M. 1999. Coordinating restraint: Democratization, fiscal policy and money creation in Latin America. *Political Research Quarterly* 52:729–51.

Saideman, Stephen M., David J. Lanoue, Michael Campenni, and Samuel Stanton. 2002. Democratization, political institutions, and ethnic conflict: A pooled time-series analysis, 1985–1998. *Comparative Political Studies* 35:103–29.

Shellman, Stephen M. 2004. Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis* 12:97–104.

Stimson, James A. 1985. Regression in time and space: A statistical essay. *American Journal of Political Science* 29:914–47.

Wawro, Gregory. 2002. Estimating dynamic panel models in political science. *Political Analysis* 10:25–48.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data.* Cambridge, MA: MIT Press.

Zahariadis, Nikolas. 2001. Asset specificity and state subsidies in industrialized countries. *International Studies Quarterly* 45:603–16.