A Lot More to Do:
The Promise and Peril of Panel Data in Political Science

Corresponding Author:
Sven E. Wilson, PhD
Assistant Professor
Department of Political Science
Brigham Young University
732 SWKT
Provo, Utah 84602
Phone: 1-801-422-9018
Email: Sven_Wilson@byu.edu

Daniel M. Butler
Department of Political Science
Encina Hall West, Rm. 100
Stanford University
Stanford, California 94305
Phone: 1-650-497-7462
Email: Daniel_Butler@stanford.edu

February 16, 2004

Abstract

In 1995, Beck and Katz (B&K) instructed readers of *APSR* on "What to do (and not to do) with time-series, cross-section data." Even though this influential paper largely ignored the extensive literature on panel data methods, the simple B&K prescriptions rapidly 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 135 papers we review show a widespread failure to diagnose and treat common problems of time-series, cross-section (TSCS) data (such as unit heterogeneity), to consider alternative dynamic specifications, to account for autocorrelation, and to acknowledge the unpleasant fact that reliable small-sample methods of estimating dynamic models with unit heterogeneity (which characterizes *most* TSCS analysis in political science) *do not yet exist*. Furthermore, we replicate eight papers in prominent journals and find that simple alternative specifications often lead to drastically different conclusions. We summarize many of the statistical issues relative to TSCS data and show that there is far more to do with TSCS data than B&K led us to believe.

* 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.

1. Introduction

In 1995, Nathanial Beck and Jonathan Katz (B&K) published an article in the *American Political Science Review* entitled "What to do (and not to do) with time-series, cross-section data." The main thrust of the paper was a stinging and effective critique of Parks' (1967) FGLS method, which was the prevailing orthodoxy in comparative politics research at the time for models employing time-series cross-section (TSCS)[1] data. In addition to obliterating the use of Parks' method in political science research, Beck and Katz (B&K) offered some guidelines on how to treat TSCS data, including the introduction of panel-corrected standard errors (PCSEs).

The rapid rate at which political scientists have adopted their prescriptions has few parallels in social science research. Within a short time, commercial statistical packages developed canned routines to estimate panel PCSEs and, as of May 31, 2003, there had been 170 citations in the political science literature to the 1995 paper with probably hundreds more in press or in progress.[2] This new orthodoxy was conceived with little discussion of or even reference to the large body of literature on TSCS methods that had been established decades before 1995. This oversight led many researchers to implement dutifully the B&K method as if it were the comprehensive guide suggested by the paper's title.

Many scholars approach TSCS data with an insufficient appreciation of the challenges inherent in applying regression methods in this context. These challenges can be categorized as follows. First, all the normal problems that arise with cross-sectional analysis persist, but now the fundamental assumption that the observations are independent is violated because there are repeated observations on the same analytical unit (usually this consists of yearly observations on the same set of countries). This hierarchal data structure, thus, violates the assumption that observations are independent from one another, a

[1] Beck (2001) articulates a distinction between panel data and TSCS data based on the notion that TSCS data has a fixed, non-sampled N (number of unites), whereas panel data, according to his definition, has a short time horizon and a large number of randomly sampled units. Although we think that this definition of panel data is too restrictive, we follow the same convention. The important point is that TSCS data follows a hierarchal data structure, and most of the issues relevant to hierarchal data apply to both types of data sets—those with a small, fixed N and with a large, sampled N.

[2] This is based on an electronic search of the Social Science Citation Index (SSCI). There were an additional 82 citations in the SSCI outside of political science.

problem not corrected by PCSEs.[3] Second, all the complexities of time-series analysis are present. Myriad plausible dynamic specifications exist for estimating time-series data, but little of the published TSCS political science literature approaches the question of dynamics seriously. Finally, TSCS data are almost always small in either number of units (*N*) or length of time (*T*)—usually in both—which compounds all the problems just mentioned. Consequently, estimation techniques that rely on large sample properties to correct for violation of the basic OLS assumptions cannot be relied upon in most instances, and small sample properties associated with different methods of dealing with dynamic panel data models are largely unexplored.

Our intent is not to explore new methodological territory nor to establish a new, one-size fits all approach for doing TSCS analysis (though we do try to provide a minimal survey of the panel data methods B&K (1995) neglected). But ours is a tale that deserves telling not just for its methodological advice, but because it contains a moral that transcends any of our specific critiques. 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 researcher have ignored their advice (they have not); rather, many researchers have followed the prescriptions far too exactly and have been blind to theory, data characteristics, a variety of specification issues, alternative models, appropriate diagnostics, and long-established pitfalls of regression analysis. In short, this is a tale not only about methodologists giving advice, but also about those who eagerly follow such advice.

2. TSCS Models: Important Things They Didn't Tell You.

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 B&K method—as it is practiced in the literature—consists of three essential components:

1) Pool the data from different countries into one dataset and apply OLS

[3] PCSEs do, however, account for some types of correlation between the error terms.

2) Adjust for autocorrelation by adding a lagged dependent variable to the model.
3) Calculate PCSEs.

Beck and Katz make an important contribution with the introduction of PCSEs, which consist of a straightforward and reasonable application of the well-known formula for estimating the standard errors of OLS regression coefficients when the error-distribution is non-spherical.[4] However, appropriate TSCS methodology addresses problems more fundamental than heteroskedastic or autocorrelated error terms, though those are important second-order concerns. More important is the specification of the estimating equations themselves. To this question, B&K offered virtually no guidance in their 1995 paper and only a little more in their 1996 paper on dynamics. In the sections that follow we attempt to highlight a few of the important issues largely neglected in both the B&K papers and, consequently, in the bulk of the TSCS literature in political science. Because most of the applications of the B&K method are in an international context where countries are the primary unit of observation, in what follows we follow the rhetoric of referring to units as "countries." However, the issues we raise are equally relevant to other units of analysis such as states, counties or other divisions within countries and to observational units not characterized by geography.

While we cannot address all the issues involved with TSCS analysis, we structure our critique around two broad classes of methodological misdeeds that have been prevalent in the literature. The first is the failure to account for unit heterogeneity or, in other words, the failure to account for unobserved local factors. Indeed many researchers fail to realize that TSCS data provides a valuable means of mitigating omitted variable bias, which is *the* fundamental and pernicious problem in regression analysis. We highlight the importance of heterogeneity and the simple models that can be used to explore and correct it.[5] The second misdeed is lack of attention to dynamic issues beyond the simple lagged

[4] Given a non-spherical covariance structure expressed by $E(uu'|X)=\sigma^2\Omega$, the variance of the OLS estimator is: $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}(X'\Omega X)(X'X)^{-1}$. PSCEs are obtained by treating Ω is an $NTxNT$ block diagonal matrix with $\hat{\Sigma}$ along the diagonal where $\hat{\Sigma}_{i,j} = \frac{\sum_{t=1}^{T} e_{i,t}e_{j,t}}{T}$

[5] Interestingly, Beck (2001) refers to models of heterogeneity as "modern" approaches, even though heterogeneity models were standard in textbooks long before 1995.

dependent variable approach encouraged by B&K. Specifying dynamics in time series data is among the most challenging tasks applied researchers face, but our review finds precious little attention paid to these crucial issues, even to the limited extent discussed in B&K (1996).

2.2 Accounting for unit heterogeneity with hierarchal models

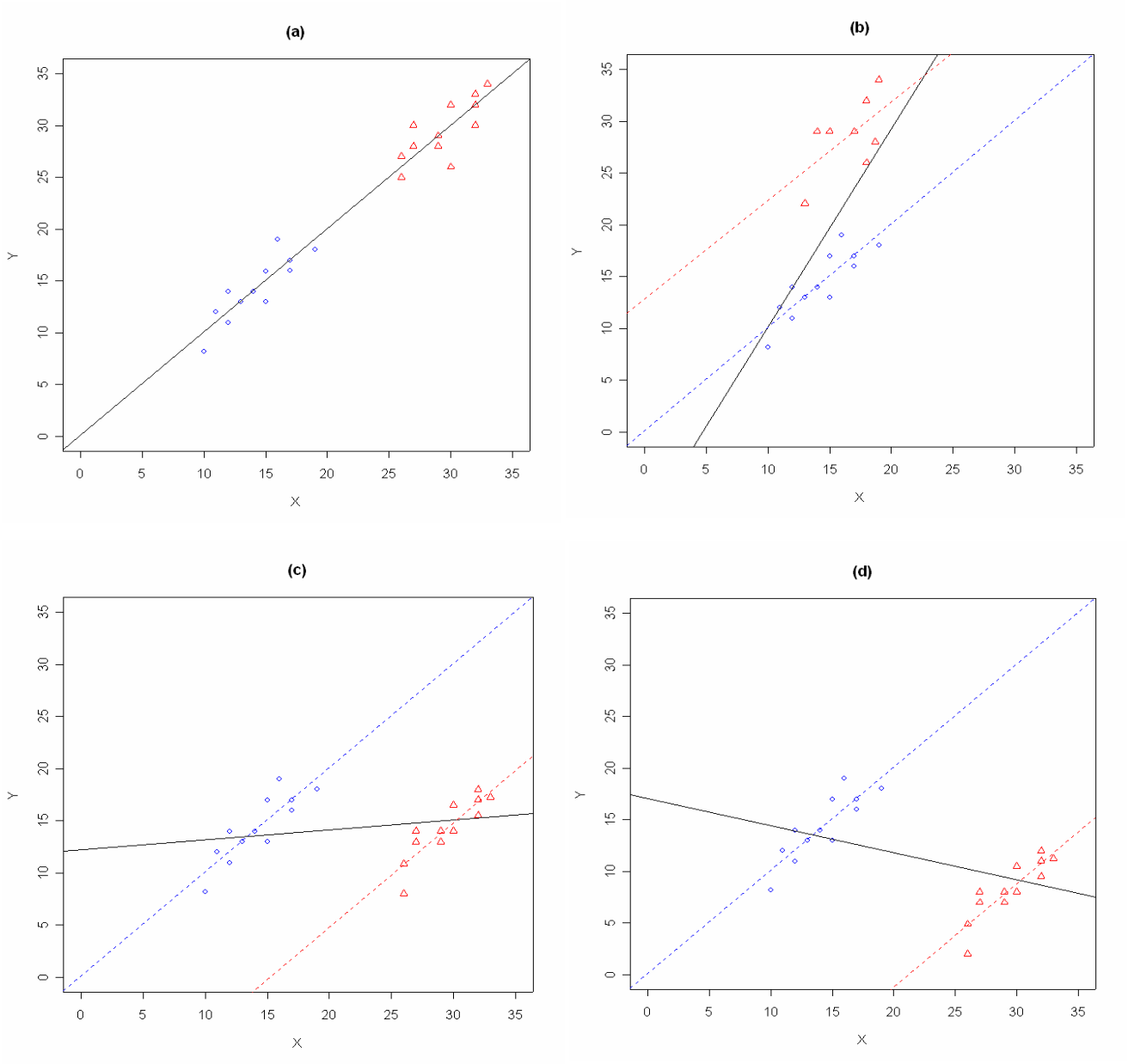
Unit heterogeneity means that countries differ in ways not explained by observed independent variables. In other words, potentially important local factors are unobservable to the researcher, which results in omitted variable bias. Standard OLS regression does not offer a solution to omitted variable bias, and PCSEs are also not a solution. In their 1995 paper, B&K led readers to believe that the most fundamental problem with TSCS data is specifying the error structure of the model. Far more essential is the question of how to specify the basic estimating equation and the related question of how to treat the (possibly heterogeneous) data from different countries.

When researchers use OLS on data pooled from different countries, they implicitly assume that unobserved local factors do not exist. Figure 1 illustrates the potentially severe consequences that can result from using OLS on pooled data—even if the slopes and error distributions are assumed to be identical across countries. In this example, each of the two countries has data characterized by the simple linear regression model $Y=\alpha + \beta X + u$, where u is a random error term. In all four cases, the error distributions are identical and each country has the same slope coefficient, $\beta=1$. Case (a) is the situation assumed by the simple pooling model of B&K. Case (b), on the other hand, shows a situation where the intercept for country 1 is lower than the intercept for country 2, which results in an overestimation of β when the data are pooled. Similarly, case (c) shows the situation where the pooled regression underestimates the slope, and case (d) illustrates that pooling can even cause the estimate to be of the wrong sign. In sum, even though all countries may be characterized by the same slope, variation in the intercepts and mean levels of the independent variables will result in incorrect parameter estimates.

2.2.1. Basic hierarchal models

Missing variables constitute a fundamental problem, but TSCS data provide a straightforward method for chipping away at that problem. Because we have ample reasons to think that some of the

Figure 1: The Consequences of Pooling Data



(a) Pooled regression correctly estimates slope; (b) pooled regression overestimates slope; (c) pooled regression underestimates slope; (d) pooled regression estimates incorrect sign for slope

omitted variables are country-specific (as opposed to randomly distributed within and across countries, as assumed by pooling), we can begin to address the omitted variable problem by explicitly modeling the differences between countries. In other words, we allow for a two-level hierarchy in the specification of the equations—some variables vary from year to year *within* countries, but other variables are constant across years and vary only *between* countries.

A model with varying intercepts is the most basic hierarchal specification. This can be expressed as a simple deviation from the standard linear model as follows:

$$Y_{it} = \beta X_{it} + \alpha_i + u_{it} \qquad (1)$$

In this case, *i* refers to the group or unit (a country, in our case), and *t* refers to the individual observation (the year) within the group.[6] The key difference in this model from standard OLS is that the intercept, α_i, is assumed to vary across countries but be constant within a country. Given this specification, researchers must ask whether the α_i terms should be treated as fixed parameters to be estimated or as random variables. The random effects model[7] (REM) model assumes that these are random draws from the same distribution and therefore part of a composite error term $e_{it} = \alpha_i + u_{it}$. The fixed effects model (FEM) assumes the α_i are fixed, estimable parameters. In other words, the errors in estimation shown in Figure 1 are corrected in the FEM by letting each country have its own intercept.

Although the differences between the FEM and REM seem benign at first glance, they are actually very different approaches to estimation. If the α_i can be considered fixed parameters to estimate, then the FEM has all the properties of the simple OLS model (estimates that are unbiased, efficient, and asymptotically normal, for instance). The model can be estimated by OLS by subtracting the country-specific mean from each observation and estimating the following equation:

$$Y_{it} - \overline{Y}_i = \beta(X_{it} - \overline{X}_i) + v_{it}. \qquad (2)$$

[6] In the case of TSCS data, the individual observations refer to observations on the same analytical unit at different points in time, but the time subscript could be generalized to other concepts (such as geographical sub-units).
[7] In econometrics, the random intercept model is called the "random-effects" model, a convention we use here, but in other fields (such as psychometrics) the term "random-effects" is used more generally, and the model discussed above is just the simplest case (sometimes called the "variable intercept model") of a class of more complex models.

Although it may not be obvious, the exactly same results can be obtained simply by including a dummy variable for each country or unit (and deleting the intercept). For this reason the FEM is sometimes called the Least Squares Dummy Variable (LDSV) model. Most statistical packages include routines to easily estimate the FEM. Dummy variables can also be done "by hand." This is useful because the signs and magnitudes of the coefficients on the dummy variables may be substantively important, and the FEM detects idiosyncratic variation across countries[8] (how much, for example, does France differ from Germany in ways not explained by the other independent variables?).

There are two main drawbacks to the FEM, however. First, because the model introduces *N* new parameters into the model, it can chew up degrees of freedom. This will cause less precise estimates of the other parameters in the model. For this reason, the researcher should always test whether the fixed effects are jointly significant. If they are not, then they can be deleted in favor of the simple pooled OLS model of B&K. It is also possible to delete subsets of the fixed effects parameters (which implies that there are a few idiosyncratic countries that need to be controlled for). However, dropping variables in an *ad hoc*, atheoretical fashion should be avoided. The second main drawback is that variables that are fixed over time cannot be included in the FEM, and variables that change only slowly will likely have very large standard errors. Issues of inference for these types of variables are discussed in Section 2.5. But even if the FEM is ultimately not used to make inferences about the parameters, it will always be a useful diagnostic tool—it tells the researcher about the potential importance of unobserved heterogeneity, and it can identify those units (countries) that are particularly influential.

Because it is a more efficient (only one additional parameter need be estimated) than the FEM, the REM may seem attractive to researchers. But we should not forget the central contribution of B&K: FGLS (Feasible Generalized Least Squares) techniques can lead to seriously incorrect standard errors in small samples. The REM requires a FGLS[9] approach that is justified only asymptotically. In general, the

[8] In fact, detecting which countries are idiosyncratic may be the central research objective in some cases.
[9] FGLS approaches involve making assumptions about the covariance structure of the error terms and then transforming the data before running OLS. B&K point out that these transformations can sometimes involve exceedingly poor estimates of the covariance matrix. In the REM, the transformation assumes a covariance structure that accounts for the fact that the α_i are constant within units, but follow a random distribution across units.

small-sample properties of FGLS estimators are not well known. Furthermore, the REM requires that the unobserved effects are uncorrelated with the other explanatory variables in the model. This assumption is often violated; indeed, a main reason for worrying about heterogeneity in the first places is that we are worried about the bias caused by such a correlation. Finally, the REM only makes theoretical sense if we can conceive of the intercepts as random draws from a larger universe. When the units of analysis are countries, the random variable assumption is difficult to rationalize. We do not see the REM as a generally useful or appropriate approach to accounting for heterogeneity in TSCS data.[10]

Although less used, a third variant is the between effects model (BEM). In this method, time is completely removed from the data and the following regression is run:

$$\overline{Y}_i = \beta \overline{X}_i + u_i . \qquad (3)$$

A major shortcoming of the BEM is that it completely assumes away any important dynamic effects, so cases in which variation over time matters, the BEM is not useful. Furthermore, the BEM reduces the number of observations from *NT* to *N*, which drives us back to the small-*N* problem where we do not have enough observations to do us any good. But if what the researcher is observing is essentially repeated observations of the same phenomena or if the primary variables of interest do not change significantly over time, the BEM is the most conservative and appropriate approach (more on this issue in Section 2.5).

Before moving on, it is useful at this point to distinguish unit heterogeneity from panel heteroskedasticity. Panel heteroskedasticity occurs when the variance of the errors (assumed to be constant *within* countries in the B&K framework) differs across units. For instance, the level of exports in one country might be more variable than another. Although PCSEs account for contemporaneous correlation of the error terms, which would allow for shocks to the level of exports to be correlated within groups of countries (the European Union, for example), the B&K framework still assumes that the error terms have a zero conditional mean over all subsets of the data.[11] Unit heterogeneity, on the other hand,

[10] A test by Hausman (1979) can be used to test the random effects assumption that the unit effects are uncorrelated with the error term. If the null hypothesis is not rejected, the REM will be a more efficient estimator than the FEM, but the other problems with the REM still exist.

[11] More technically, that $E(u_{ij}|X_{ij})=0$ for all *i* and *j*.

occurs when the α_i terms (the unit effects) are nonzero. This implies that subsets of the data (namely countries) do not have a conditional mean of zero because the true error term is the composite error, e_{it}, = $\alpha_i + u_{it}$, as defined above.

To summarize, when a natural grouping of the data exists (in other words, when the data is hierarchal), such as observations organized by country, the researcher can exploit knowledge about the hierarchal structure of the data to improve the estimation. Viewed in the appropriate light, TSCS data is a blessing because it allows researchers to *model* the data dependence, rather than simply suffer from it. Later (in Section 3.2) we show the sensitivity of several published results to models that account for heterogeneity.

2.2.2. Generalized hierarchal models

In many respects, the basic hierarchal models are only starting point. It is the simplest version of a more general class of panel data models that allow intercepts and slope coefficients to vary across countries (or other types of units) as follows:

$$Y_{it} = \beta_i X_{it} + \alpha_i + u_{it}. \quad (4)$$

From a theoretical standpoint, allowing the slope coefficients to vary is highly intuitive. However, introducing this additional random variation and then modeling it places additional strains on small data sets. We do not, however, want to dissuade researchers from exploring random coefficient models or models that further generalize (4) by adding additional levels of hierarchy (such as countries existing as part of regions, where there are regional-level effects).

Because of space considerations we do not undertake an analysis of these more general models here. It is worth noting however, that these models go by a variety of names: hierarchal models, random-effects models, random coefficient models, correlated data models, mixed models, or multi-level models. Bayesian versions of these models exist as well and should be explored. Recent important

methodological surveys apply standard tools from statistics and other disciplines to political science.[12] We also ignore here other important issues that arise in cross-sectional analysis (and, hence, in TSCS analysis). Chief among these may be the problem of endogeneity, meaning that variables included in the regression are correlated with the error terms. Controlling for heterogeneity through the FEM accounts for the endogeneity due to unobserved, unit-level heterogeneity, but there are other sources of endogeneity as well. Instrumental variables estimation is the usual approach to tackle this problem and can be implemented in a TSCS context.

2.3 Taking time seriously

B&K followed their 1995 study with a more substantive (but much less cited) analysis of dynamic issues related to TSCS data in a 1996 article in *Political Analysis*. Using a simulation analysis, they conclude with the prescriptions of their earlier analysis but add two important caveats: 1) the estimation should be followed by a test for serial independence of the error terms and 2) testing to see if a more general dynamic model is appropriate. B&K suggest the lagged dependent variable (LDV) model, in part, because it will cause researchers "to think about the dynamics of their model" (p. 12). As our review of the literature (Section 3.1) suggests, however, the 1995 paper by B&K model has probably caused researchers to think *less* about dynamics because it leads readers to think that the LDV model with PCSEs already accounts for dynamics. In any case, a majority of researchers are not exploring the issue of dynamic specification, at least in the published literature (which we demonstrate in Section 3.1).

While published papers in political science pay much attention to theories about why X affects Y (or why Z should not affect Y), there is less often theoretical attention to how X should affect Y over time. And to say that "the past matters" does not tell us much concerning *how* it matters. Even when researchers employ a theoretical justification for a particular specification, it is often true that the same theory can be used to justify another specification as well.

[12] Important examples are Western (1998), who highlights Bayesian hierarchal methods; Zorn (2001), who employs a generalized estimating equation approach; and Steenbergen and Jones (2002).

An enormous variety of dynamic models have been explored in the literature, and we can discuss but a few here. In what follows we employ some limiting assumptions. First, we limit our discussion to models that have up to one lag in the dependent and independent variables. Second, we drop the country index, *i,* and all references to lags should be interpreted to be within-panel time lags. Finally, the independent variables are represented by a single variable *X*, though all of the results apply to the case where X is considered a vector containing multiple independent variables.

The first two models we consider are the simple static model and the static model with an autoregressive error term. These are stated as follows:

Static Model:

$$Y_t = \alpha + \beta_0 X_t + u_t . \qquad (5)$$

AR(1) Model:

$$Y_t = \alpha + \beta_0 X_t + u_t ; \qquad (6)$$

$$u_t = \rho u_{t-1} + e_t \qquad (7)$$

These two models are the workhorses of regression analysis with time-series data, but they essentially provide no dynamics (no lagged values of *X* or *Y* are included). Distributed lag (DL) models, on the other hand, provide a means of including dynamic relationships into the model:

DL(1) Model:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + u_t \qquad (8)$$

A central task in estimating distributed lag models is estimating the number of lags to include in the model (again, we are assuming that higher-order lags are not present). This has both empirical importance for obtaining unbiased estimates and theoretical value.[13] An advantage of the DL model is that no particular statistical complications arise in adding lagged values of the independent variables

[13] Virtually any introductory text in Econometrics will have a discussion of distributed lag models, as well as non-linear distributed lag models. We recommend Gujaratti (1995) for an excellent discussion.

model as long as the X values are uncorrelated with the error term. Should the errors be autocorrelated or heteroskedastic, they can be dealt with as they would be in the static model.

Statistical complications increase enormously, however, if the researcher places a lagged dependent variable (LDV) into the model, which is what B&K recommend:

LDV Model:

$$Y_t = \alpha + \beta_0 X_t + \gamma_1 Y_{t-1} + u_t . \qquad (9)$$

This is the model suggested by B&K and used by the majority of researchers in recent years in political science. Although it is a true dynamic model with a natural intuitive appeal it poses serious problems. First, if the error terms in (9) are autocorrelated, OLS estimate of all the LDV model's paratmeters will be biased and inconsistent.[14] Ironically, B&K suggest the LDV model as a *solution* to the autocorrelation problem of the AR(1) model even though the LDV may actually *introduce* bias that is not present in the AR(1) model. Second, Achen (2000) has shown that the LDV model can lead to highly spurious results. He demonstrates that the estimated coefficient γ_1 can be highly significant and the R^2 can be very high, even if there is no impact whatsoever of the LDV in the true model (in other words, when $\gamma_1=0$). More generally, the LDV model essentially uses the dependent variable to explain itself, which is something that should always give researchers some pause. Finally, it has been known for some time that including an LDV when there is unit heterogeneity will lead to biased and inconsistent estimates, which we discuss further in Section 2.4. It is, of course, possible that the LDV model is either theoretically motivated or is the model that best fits the data. In this case, the LDV should be included, but the issues just discussed still need to be addressed.

The model that contains both the DL(1) and the LDV(1) as special cases is the auto-regressive, distributed-lag model:

ARDL(1,1) Model:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \gamma_1 Y_{t-1} + u_t \qquad (10)$$

[14] The formula for this bias is well known and can be found, for instance, in Greene (2000, pp. 534-535).

The ARDL(1,1) model also includes other models as special cases. These include several autoregressive models including the Koyck model, the adaptive expectations model and the partial adjustment model.[15] Furthermore, the AR(1) model above can (after a little algebra) be written as the following ARDL(1,1) model:

$$Y_t = \alpha(1-\rho) + \beta X_t - \rho\beta X_{t-1} + \rho Y_{t-1} + e_{i,t}, \qquad (11)$$

In other words, the ARDL model can be used to correct for first-order autocorrelation in the same way as the AR(1) model, though the coefficients on the lagged regressors of the ARDL model need to be interpreted differently in this case. A particularly unfortunate consequence of B&K (1995) is that readers were led to believe that the LDV model is the appropriate solution to autocorrelation.

Given the generality of (10), the preference given by B&K to the LDV model is somewhat perplexing. They use an ARDL(1,1) model to generate the data for their Monte Carlo analysis in B&K (1996), but then argue that while neither the LDV approach nor the simple Prais-Winston transformation is the same as estimating (10), they claim that in practice either method is "good enough" (p. 9). It is not clear why they do not advocate simply estimating (10) directly (though this requires a more nuanced interpretation of the independent variables since they appear in both current and in lagged form). We argue that estimating the ARDL model is an extremely useful diagnostic tool because it nests these other models as special cases and standard *F*-tests can be used to test the significance of the variables in the model (dummy variables for each country can also be included as part of this diagnostic testing).

The final dynamic model we will consider is the first difference (FD) model:

FD Model:

$$Y_t - Y_{t-1} = \beta_0(X_t - X_{t-1}) + u_t \qquad (12)$$

FD models assume that the *change* in *X* (rather than the absolute level of X) causes *Y* to change over time. The FD model also provides an alternative solution to the problem of local factors, since the intercept term, which might vary across countries, is effectively "netted out" of the equation. As B&K (1996) note,

[15] See, for instance, Gujarati (1995, p. 602)

however, the first difference model captures only the short run dynamics,[16] But this criticism is equally true of the LDV model and other dynamic models discussed here.

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 should be considered more plausible than any of the other dynamic models. It is important to note that *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. Without further theoretical justification, each of these models seems as plausible as the next, but they differ in their levels of generality and robustness, and more general models (such as higher order lags) could be explored.

2.4 Combining unit heterogeneity and dynamics: Bias in dynamic panel models

A "dynamic panel model" is a model characterized by both unit effects (unit heterogeneity) and dynamics (often the presence of an LDV). However, it has been known at least since the Monte Carlo studies of Nerlove (1967, 1971) that introducing an LDV into the FEM results in biased estimates. The estimates are also inconsistent (in *N*) as long as *T* is finite, and Nickell (1981) has derived the exact formula for the bias under OLS. To see why this bias exists, recall that the FEM can be obtained by "de-meaning" the data as follows:

$$(Y_{it} - \bar{Y}_i) = \beta_0 (X_{ij} - \bar{X}_i) + \gamma_1 (Y_{it-1} - \overline{Y}_{i,-1}) + (u_{it} - \overline{u}_i). \qquad (13)$$

When T is small, the mean residual *within a country*, $\overline{u}_i$, can be large and have a high variance across countries.[17] Furthermore, when a country has unusually (meaning not explained by X) high values of Y, it will also have unusually high values of both the lagged value of Y and the residual, which means that $\overline{Y}_{i,-1}$ will be correlated with $\overline{u}_i$. This, in turn, implies that the term $(Y_{it-1} - \overline{Y}_{i,-1})$ is correlated with the

[16] Another type of dynamic model mentioned by B&K (1996), but not explored here, is the "error-correction" model.

[17] The mean residual for the sample as a whole will be zero by construction, but of course this does not hold for the mean of any subset of the data.

error term $(u_{it} - \overline{u}_i)$, which causes *all* the parameter estimates of the model to be biased. Naturally, as T grows, the bias will disappear because the variance of $\overline{u}_i$ across countries will go to zero.

Several estimators have been proposed to deal with this issue including a GMM-based estimator of Arellano and Bond (1991), a "corrected" LSDV estimator of Kiviet (1995), a "nearly unbiased" estimator of Carree (2001), and a maximum likelihood estimator by Hsiao, Pesaran and Tahmiscioglu (2002).[18] In general, however, the properties of these and other estimators are based on asymptotic results in N. The problem is that in TSCS data used by political scientists, N is usually very small. Unfortunately, very little work has been done to understand the properties of these new approaches when applied to the types of data sets political scientists are likely to use. An initial encouraging result is that Judson and Owen (1999) have found that the bias on the X variables is quite small in Monte Carlo analysis, though the bias on the coefficient of the lagged dependent variable can be substantial. They also find that the corrected-LSDV model (which is a modified version of the FEM) performs the best in their simulations. However, far more Monte Carlo results are need to determine which of these new estimators are more reliable and under what conditions.

Monte Carlo analysis by Kristensen and Wawro (2003) illustrates the problems that arise with the B&K method when LDVs are included in a model where unit effects exist. Naturally, the problems with the LDV increase as the correlation between the unit effects and the observed variables increases. As far as we know, what still needs to be demonstrated is that bias will be associated with data generating processes that are "near" the FEM, such as a model with time-invariant or trended explanatory variables. We suspect that types of data processes where OLS is biased in the case of an LDV is much more general than the simple FEM.[19] Similarly, dynamic structures that are "near" the LDV, such as model with two lags, are also likely to generate biased estimates with OLS.

---

[18] The literature on this topic is large. Wawro (2002) gives a review of some of these estimators and their application to political science. See also Bun and Carree (2002), Ahn and Schmidt (1995), Arellano (1989), Bun and Kiviet (2001), Chamberlain (1980), Hausman and Tayler (1981), Heckman (1981), Jeane and Runkle (1992), and Ziliak (1997).

[19] For the intuition behind this claim, note that the FEM can be estimated by incorporating dummy variables for each country. Now, think of generalizing the FEM changing some of the 1s and 0s to slightly different values. Clearly OLS in this slightly generalized model will be biased as well.

As a final word of warning, we note that simply deleting either the LDV or the fixed effects will not eliminate the bias and may actually make it worse. Unfortunately, none of the estimators developed thus far have shown to be a reliable solution in small samples. Simply put, we do not know enough yet about the small sample properties of dynamic panel models in small TSCS data sets to make reliable inferences from our estimates. Since many TSCS data sets in political science are going to have significant unit effects (meaning some country-specific effects are unobserved) and important dynamic relationships (the value of *Y* today depends in some way on past values of *Y*), most of the estimates obtained with the B&K—or any other—method are problematic. We see no benefit from sugarcoating this unpleasant fact.

2.5. Pooling and the problem of sluggish variables

At its heart, the B&K method is a valiant attempt to turn small samples into large ones. In Figure 1, we showed how pooled-OLS can have serious consequences in the presence of unobserved local factors (a theoretical point that we confirm empirically in Section 3.2). The assumption of the B&K method is that data from different countries can be combined and analyzed together. Rather than simply employing the pooled-OLS approach of B&K, hierarchal models (such as the FEM) can be used to diagnose and correct for different forms of cross-country heterogeneity, as we have shown. Alternatively, country-by-country regressions can be used to compare the effects of explanatory variables across different countries.

In using TSCS data, a fundamental question is whether the repeated observations within a country can be considered as legitimate. Researchers must be very cautious to avoid confusing observations with cases, as Kittel (1999) points out. For example, consider investigating the effect of regime type on the provision of public goods with data on 20 countries for 20 years, and suppose regime type for each country does not change over the 20 year period. Suppose we were to increase our rate of observation to monthly. Would we then have 4,800 observations? Surely such an approach would constitute an artificial inflation of sample size. But why is the smaller sample of 400 observations any different?

In TSCS data there are two sources of variation: cross-sectional and dynamic. When significant variation exists across time—in both the dependent and independent variables—each year constitutes a legitimate observation. However, if little variation occurs across time, then what justifies treating each year of data as a unique observation? Since in this case it is only the cross-sectional variation driving the results, the BEM (which regresses mean values of Y on mean values of X) seems to be much more legitimate than either pooled-OLS or FEM, though BEM will seldom result in statistically significant effects.

The conceptually more difficult problem occurs when the dependent variable varies significantly over time, but a key variable of interest does not. 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 to the time-invariant and slowly-moving variables as "sluggish," and the FEM model is a poor context with which to analyze sluggish variables. As Beck 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 (2001).

We definitely agree that unit effects "soak up" the explanatory power of sluggish variables, but in our view this—to the extent that following conservative norms of inference is desirable—is a good thing, not a "cost." Controlling for heterogeneity raises the bar for confirming our theories. Certainly the researcher would always want to run the model with and without unit effects because both specifications yield useful, substantive information. And the researcher might be fortunate in that the variables of interest are robust to the inclusion of unit effects. We find in the next section that some empirical findings are actually *strengthened* when unit effects are added to the model.

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 FEM is to

fundamentally misunderstand the role of statistical analysis in theory evaluation. If we *knew* the true model 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 an alternative explanation (almost always a plausible one), that unobserved local factors drive, at least in part, the cross-country variation in the dependent variable.

Some scholars clearly understand the role that unit effects play in cross-national comparisons. Rueda and Pontusson (2001), for instance, argue:

> "Scholars engaged in cross-national comparison sometimes eschew the use of country dummies on the grounds that they simply tell us that countries are different, when the interesting question is how or why they are different. Yet there is every reason to suspect that outcomes, such as ours are influenced by country-specific historical or cultural factors, which cannot be measured on a cross-national basis…" (p. 369).

All the FEM tells us is that France is systematically different than Germany, but the researcher often wants to know *why* France is different than Germany. If the researcher has a strong theory of this difference that is built upon sluggish variables (such as regime type) it will be hard to estimate this effect with the FEM. In response we note that just because the researcher wants to test a good theory does not mean it is possible to do so with the data at hand. We certainly endorse the effort to *explain* cross-country differences—rather than merely control for them—but the problem with the pooled-OLS solution is that it implicitly assumes the researcher has 400 observations on which to test her theory. In fact, she does not have 20 Germanys, 20 Frances, and 20 Norways; she has only one of each. There are only 20 cases, not 400.

3. Does Method Matter?

3.1 The pied pipers of panel data: A methodological review

Of those papers that have cited B&K (1995), we identified 135 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 May 31, 2003.

We do not analyze non-linear 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.[20] 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 LDV model. Our analysis gives each paper a strong benefit of the 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 40.7% of the studies reviewed considered unit-effects. This number is somewhat encouraging given the lack of attention to heterogeneity in the 1995 paper. Furthermore, 85.5% 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.

On the issue of dynamics we found that that a surprising 32.6% of studies have models with *no* dynamics. Of those 44 studies, 39 (88.6%) provide no justification for why they are ignoring dynamic issues. The most common model, not surprisingly, is the LDV model encouraged by B&K. We find that 43.7% of studies report the LDV model without considering any alternative specifications. Of those who use the LDV model, only 37.3% 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 20 studies use the LDV model for theoretical reasons. Only 32 (23.7%) of the studies consider alternative dynamic models, though several of these rely predominantly on the LDV model for their main inferences and only 37.5% of them test for autocorrelation.

Any careful study using TSCS should consider unit heterogeneity and alternative dynamic specifications and test for autocorrelation. We conclude that only 14 of the 135 (10.4%) studies consider

[20] The complete data for this review can be found in Appendix A, which is available at http://address hidden for purposes of review.

Table 1: A Summary of Key Methodological Issues in Published TSCS studies in Political Science

135: Studies Reviewed

Panel Heterogeneity

55 (40.7%): Number that consider fixed-effects

- **8** (9.1%) Number which test and reject fixed-effects before excluding
- **47** (34.8%) Number which report fixed-effects

80 (59.3%): Number that do not consider fixed-effects

Dynamics

44 (32.6%): Number that use models with no dynamics

- **5** (11.4%): Justification provided for not using dynamics
- **39** (88.6%): No discussion provided

59 (43.7%): Number that consider and use only the LDV model

- **22** (37.3%): Number that test for autocorrelation

Reason(s) given for including the LDV

- **20** (33.9%): Theoretical
- **36** (61.0%): To correct for autocorrelation
- **35** (59.3%): Recommended by B&K

32 (23.7%) Number that use or consider alternative dynamics

- **12** (37.5%): Number that test for autocorrelation

Notes: This is based upon articles found doing a search for articles citing Beck and Katz (1995 or 1996) on the Social Sciences Citation Index on May 31, 2003. This list only represents those articles found in that search that used linear models with time-series cross-section data. Articles using other methods were not included in this table.

both heterogeneity and alternative dynamics. Of those, only 6 test for autocorrelation. Many studies do have thoughtful analysis of methods, but almost all are incomplete in important ways. 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. In numerous cases, it is clear that researchers are using B&K (1995) just as its title suggests—as an 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 that a reliable method for estimating panel data models in small samples has not yet been developed, as we discussed in Section 2.4.

3.2 Robustness

To put our criticisms to the test, we chose 8 published studies and re-analyzed the data incorporating unit effects and alternative dynamic specifications. We test first for robustness with respect to unit heterogeneity by comparing estimates of OLS estimates to estimates from the FEM. We then turn to the issue of dynamics by re-estimating the models using the six models discussed in Section 2.3.

We did not choose studies for analysis randomly, nor do we make broad claims about their representativeness. Of the 135 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 8 studies for replication. We were somewhat biased towards those studies which had data available on-line, 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 here that these studies are not the worst offenders of the problems that we have discussed here. For example, one study tests an alternative dynamic

Table 2: Relationships Tested in Sensitivity Analysis

Article	Page #	Table.Column	Dependent Variable	Important Independent Variable(s)
Cox et al. (*WP* , 1998)	466	1.1	Total expenditures per elector	Victory margin
Cox et al. (*WP* , 1998)	467	2.2	Voter turnout	Total expenditure, victory margin, Percent men, Percent urban, Percent of population under 15
Hood et al. (*LSQ* , 2001)	611	1.1	Unadjusted LCCR scores	GOP strength, Black electoral strength
Moene & Wallerstein (*APSR* , 2001)	869	1.2	Gov't spending on income insurance	Inequality (90/10)
Moene & Wallerstein (*APSR* , 2001)	869	1.5	Gov't spending on income insurance	Inequality (90/50), Inequality (50/10)
Pickering (*JPR* , 2002)	328	2.2	Foreign military intervention	War experience, War experience squared
Poe & Tate (*APSR* , 1994)	861	1.4	Personal integrity abuses	Democracy (Van Hanen)
Reich (*PRQ* , 1999)	743	1.3	Seniorage	First Democratic Government
Reich (*PRQ* , 1999)	743	1.4	Seniorage	Democracy for less than 10 years
Saideman et al. (*CPS* , 2002)	119	1.1	Protest	Regime type, first election, federal system, proportional democracy, enduring regime, young democracy
Zahariadas (*ISQ* , 2001)	613	2.1	Total aid	Research and development, Research and development sqaured, Job gain

specification,[21] two studies test the effect when the LDV is dropped from the model,[22] three studies give theoretical reasons for including the LDV,[23] and five studies test for serial correlation.[24] Still, seven of the articles did not consider alternative dynamic models and none of them test for unit effects.[25]

For each of the eight articles we replicated, the challenge was to pick out a set of results that were both representative of what the authors were finding, and relevant to our critiques. Our approach was to choose one or two specifications that both capture a central point of the authors' analysis and that are simple to interpret. Thus we 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 the most 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. [26]

3.2.1 Unobserved heterogeneity

Table 3 below reports the results of our replications and re-analysis when the selected models are estimated with and without fixed 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 Figure 1. We refer to a finding as strengthened or weakened if the FEM results in a change in magnitude of at least ½ of a standard error, as measured by

[21] Zahariadis (2001) included lagged independent variables along with the LDV, giving theoretical reasons for both.
[22] Cox et al. (1998), and Pickering (2002).
[23] Moene and Wallerstein (2001), Reich (1999), and Zahariadis (2001).
[24] Cox et al. (1998), Moene and Wallerstein (2001), Poe and Tate (1994), Pickering (2002), and Zahariadis (2001)
[25] While none of the authors test for fixed effects, both Reich (1999) and Moene and Wallerstein (2001) mention them. Reich mentions the fixed effects model but does not use it because "some of the measures of democratization are invariant within each panel, and would thus be perfectly collinear with country dummy variables" (741). While Reich is correct that time invariant variables will be perfectly collinear with f.e., this however should not have prevented him from testing the model on data where the measures were not time invariant (which is what we do). Further, as we have noted in the paper pooling time invariant variables without using fixed effects runs the risk of artificial sample inflation. Moene and Wallerstein go even further than Reich and note that the Devroye (2000) shows the results are not robust to a fixed effects model.
[26] Appendices B and C are found on-line at: http://address hidden for purposes of review.

Table 3: Sensitivity of Results to Unit Heterogeneity

Article	Dependent Variable	Independent Variable(s)
Estimates that Increase in Magnitude		
Cox et al. (WP, 1998)	Voter turnout	Percent men
Hood et al. (LSQ, 2001)	Unadjusted LCCR scores	*GOP strength*
Saideman et al. (CPS, 2002)	Protest	*Regime type*
Estimates that Remain Unchanged		
Cox et al. (WP, 1998)	Total expenditures per elector	*Victory margin*
Cox et al. (WP, 1998)	Voter turnout	*Victory margin*
Reich (PRQ, 1999)	Seniorage	*First democratic gov't* , Democracy for < 10 years
Saideman et al. (CPS, 2002)	Protest	First election
Estimates that Fall in Magnitude		
Cox et al. (WP, 1998)	Voter turnout	Total Expenditure
Hood et al. (LSQ, 2001)	Unadjusted LCCR scores	*Black electoral strength*
Pickering (JPR, 2002)	Foreign military intervention	War experience
Poe & Tate (APSR, 1994)	Personal integrity abuses	Democracy (Van Hanen)
Saideman et al. (CPS, 2002)	Protest	Federal system, Proportional democracy
Zahariadas (ISQ, 2001)	Total aid	R&D, R&D Squared
Estimates where Sign is Reversed		
Cox et al. (WP, 1998)	Voter turnout	Percent urban, Percent pop < 15
Moene & Wallerstein (APSR, 2001)	Gov't spending for income insurance	Inequality (90/10), *Inequality (90/50)* , Inequality (50/10)
Pickering (JPR, 2002)	Foreign military intervention	War experience squared
Saideman et al. (CPS, 2002)	Protest	Young democracy, *Enduring regime*
Zahariadas (ISQ, 2001)	Total aid	Job gain

Notes: These results represent the effect of adding unit effects (using the FEM framework) to the original model. The independent variables that are statistically significant at the .05 level after the inclusion of fixed effects are italicized. For all of the studies we used PCSE's to determine statistical significance, including Cox et al. (1998) and Zahariadas (2001) which did not use PCSE's in their initial studies. Estimates that increase (fall) in magnitude refers to those variables where the coeffiecient magnitude increased (decreased) by at least half of the initial standard earror. If the estimate did not change by more than half of the initial standard error we classified it as unchanged. The final category includes variables where the sign of the coefficient changed. Appendix B includes the regression results for the major independent variables of each study. Appendix C, contains regression results for all of the variables in these regressions.

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 GOP has pushed Democrats to the left are confirmed, but the relative importance of the GOP is increased by a factor of 3, while 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) analaysis of democracy and governmental repression are essentially robust to the inclusion of unit effects, though there are some differences in coefficient magnitude and t-statistics which 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, implying that as the number of wars (success and failures) increases, the propensity to engage in conflict increases. The FEM analysis shows just the opposite—the effect of wartime experience is now an *inverted*-U, with a maximum at .80 prior wars. This implies that each additional war (after the initial war) makes further intervention *less* likely. Similarly Moene and Wallerstein's (2001) estimates of the effect of inequality on government spending for insurance against loss of income are reversed under the FEM model, 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 non-robustness in key findings. The FEM analysis of the Saidmen et al. (2002) results show even a stronger effect of regime type than 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

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

findings from Zahariadis (2001) are weakened and the Cox, Thies, and Rosenbluth (1998) results are a mixed bag of strengthening and weakening of the key coefficients.[28]

The estimates summarized in Table 3 all 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.

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, two caveats need to be noted at the outset. First, because a finite sample will always generate non-zero correlations between the explanatory variables and non-zero regression coefficients, we will always see at least some variance across models 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

[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.

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 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 Zahariadas (2001) analysis gives estimates ranging from -2.317 to -.517. Since the PCSE is .423, this range of estimates is equivalent to 4.3 standard errors.[29] The variation is also expressed in terms of the percentage deviation from the mean estimate. The minimum and maximum deviations are reported as well as the median.

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 those estimates 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 (5 or 6) 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 that the effect of victory margin on voter turnout (Cox, Rosenbluth & Thies, 1998) has a small range of statistically significant coefficient estimates ranging between -.104 and -.082. The second category consist of "weakly robust" findings, which have at least 3 significant findings with no sign reversals (except in one instance).[30] The third category of findings includes those that are "non-robust." 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 have to do with findings that are reported in the published studies as insignificant. A "robust non-finding" occurs if the other methods all yield a small range of insignificant

[29] $((2.317-.517)/.423) \approx 4.3$.

[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, while the other models find positive and generally significant effects.

Table 4: Sensitivity of Results to Alternative Dynamic Specifications

Article	Dependent Variable	Independent Variable	Range of Estimates: Coefficients	Range of Estimates: # of std. errors	Percentage Deviations from Mean Estimate: Min.	Max.	Median	# of times variable was signifcant
Coefficients reported as statistically *signficant* using the LDV model:								
Robust Findings								
Cox et al. (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 et al. (1998)	Voter turnout	Expenditures	[-0.062, 0.065]	6.5	24%	300%	70%	4
Hood et al. (2001)	Unadjusted LCCR scores	Black electoral strength	[0.080, 2.631]	4.1	17%	91%	46%	3
Pickering (2002)	Military Intervention	War experience	[0.069, 0.164]	3.4	31%	41%	32%	3
Poe & 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
Zahariadas (2001)	Total aid	R&D	[-2.317, -0.517]	4.3	11%	74%	66%	4
Non-Robust Findings								
Cox et al. (1998)	Total expenditures	Victory margin	[-0.222, 0.017]	5.9	58%	197%	91%	5
Cox et al. (1998)	Voter turnout	Percent pop < 15	[0.419, 1.589]	20.9	2%	216%	62%	2
Cox et al. (1998)	Voter turnout	Percent men	[-0.721, 5.445]	11.3	40%	259%	140%	2
Cox et al. (1998)	Voter turnout	Percent urban	[-0.764, 0.065]	49.4	25%	410%	90%	3
Hood et al. (2001)	Unadjusted LCCR scores	GOP	[-1.391, 0.793]	9.4	316%	453%	412%	3
Moene & Wallerstein (2001)	Income insurance	Inequality (50/10)	[-4.696, 2.544]	18.1	87%	580%	197%	2
Moene & Wallerstein (2001)	Income insurance	Inequality (90/10)	[-4.442, 3.222]	24.2	197%	597%	247%	3
Pickering (2002)	Military Intervention	War experience sq.	[-0.170, 0.092]	17.9	116%	867%	153%	3
Reich (1999)	Seniorage	First democratic gov't	[0.724, 1.910]	2.6	11%	60%	26%	2
Zahariadas (2001)	Total aid	R&D Squared	[-0.021, 0.448]	3.8	24%	109%	130%	3

(cont.)

Table 4: (cont.)

Coefficients that were reported as statistically *insignificant* using LDV model:

Robust Non-Findings								
Saideman et al. (2002)	Protest	First election	[-0.198, 0.094]	2.5	78%	323%	36%	0
Zahariadas (2001)	Total aid	Job gain	[-0.017, 0.005]	2.9	48%	199%	20%	0
Weakly-Robust Non-Findings								
Moene & Wallerstein (2001)	Income insurance	Inequality (90/50)	[-23.736, -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

Notes: The B&K method uses pooled OLS with a LDV and PCSE's. It should be noted that in their initial studies, Cox et al. (1998) and Zahariadas (2001) did not use PCSE's and Saideman et al. (2002) also used the Prais-Winston 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 5 other model specifications: pooled OLS without the LDV, Prais-Winston without the LDV, lagged indepent variables (LIV's), LIV's and LDV, and first-differences. We used PCSE's in all of 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 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 .05 level. For all of the studies we used PCSE's 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 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 stastically significant in at least 3 of the specifications and had to keep the same sign. The non-robust 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 insignficant in the published reports. A robust non-finding indicates that all of the other specifications yielded a small range of insignificant effects close to zero. A weakly-robust non-finding is a result that was reported as insignficant but the range of estimates is so large that we cannot be confident that the effect is actually zero.

effects close to zero. A "weakly-robust non-finding" is a result that was reported as 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 sample sizes such as Moene and Wallerstein (2001). It is possible to have a "non-robust non-finding" 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, while 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 twelve 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 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 examine are highly contingent on the method used to obtain them. The non-robustness 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 specification. In other cases the non-robustness 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 highly frail. Of course the lesson that method matters is an old one, but it is particularly appropriate 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 non-robust 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 work in this area. Combining the theoretical potential for bias with both the sensitivity analysis we conducted above[31] 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.

Our main point, however, 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. In a recent article, for instance, Green Kim and Yoon (2001) estimated a trade model using four different specifications—1) the static model; 2) the simple FEM; 3) the LDV model; and 4) the LDV with fixed effects. Their estimates showed a dramatic and statistically significant reversal of three of the six regression coefficients (including the sacred cow of democracy) when fixed effects are added. While we are agnostic concerning which model is the best, scholars can only gain from a further analysis of this issue and efforts to understand why accounting for unobserved variables had such an impact on the results (in both the static and the LDV contexts). We suspect that if the authors had tested additional dynamic models, they would have found even additional discrepancies.

[31] We concede that it may not be valid to generalize from our non-random sample of eight studies.

Unfortunately, sensitivity analysis is not always met with gratitude. If Green, Kim and Yoon had confirmed the positive effect of democracy on trade, we wonder if the response of their critics might have been to say that "the effect of democracy on trade is very robust and persists even after controlling for unobserved variation across countries." Instead, B&K dismiss their findings out of hand. Their comment on the conflicting results is that "there is nothing in their analysis of trade….that should be seen as challenging any currently standard estimates" (p. 495)—a stunning statement, given the evidence.[32]

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*. When B&K state (in speaking of the FEM), "…to expect findings to be robust to odd specifications and or methods is a foolish expectation," (B&K, 2001) they are in effect saying that "accounting for omitted variables is foolish because your significant results might go away." Instead of treating the Green, Kim and Yoon findings as an intriguing challenge to prevailing wisdom,[33] they encourage readers to ignore them. Their failure to support high standards for theory confirmation is quite unfortunate.

4. Moving Forward

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

[32] B&K further claim that "…ignoring fixed effects simply cannot produce very biased estimates. So even if we consider…[the] possible omission of fixed effects, the consequences of this omission cannot be great" (p.492). We find this claim remarkable, in light of the fact that half the coefficients change sign in a statistically significant manner when fixed effects are added to the model—even for the LDV case.

[33] It is also a theoretically plausible finding. It could be argued that the main reason democracy promotes trade between countries is because democracy promotes economic growth, which increases the demand for trade. Representative democracies are rife with protectionists who are often highly effective at imposing trade restrictions, as both experience and public choice theorists have long noted. We also note that when fixed effects are added to the model, the effect of the GDP variable gets stronger, which is consistent with the above explanation. It is eminently plausible that in the model without fixed effects, democracy captures the effects of a host of sluggish variables that are positively correlated with trade—such as the history of trade relations between the countries or cultural, ethnic and linguistice ties—which "force" the coefficient on democracy to be positive. We hope that further research explores these kinds of questions rather than simply disregarding the findings of Green, Kim and Yoon.

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.

4.1 Specific recommendations

The most fundamental question with TSCS data is whether the data are 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 sluggish 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 sluggish variable can appear to be statistically significant. B&K have warned against the FEM because of its theoretical non-specificity, but we counter that FEM is a much more conservative approach than simple pooled-OLS because it prevents the researcher from attaching theoretical significance to variables that are merely masking the presence of unit heterogeneity. In general, it is better to lack theoretical specificity than to spuriously confirm strong theoretical claims.

In the static model, various versions of OLS and the FEM can be used to explore the issue of unit heterogeneity by estimating models with and without fixed effects, and with fixed effects alone, as well as obtaining country-by-country results to evaluate the stability of parameters across countries. Influential countries can be identified in three ways: First, the magnitude of the fixed-effect coefficients can be examined to detect countries that have a particularly large effect on *Y*. Second, deleting countries on a case-by-case basis and observing changes in regression coefficients can identify the extent to which particular countries (or groups of countries) affect the parameter estimates and t-ratios. Finally, common regression diagnostics such as DFITS or DFBETAS can be used in the FEM just as they are in the simple OLS case. In the FEM context, such techniques can identify "influential" countries as those having either

a high number of influential points or by having points with high influence values. This method can also uncover the possibility that *portions* of countries (i.e., particular years) are influential, which may be important information in its own right. In sum, the FEM model augments the standard set of regression diagnostics researchers typically use.

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 (Judson and Owen, 1999). Furthermore, simply omitting fixed effects from the dynamic models will likely cause even more serious bias.[34]

The predominance of simple pooled-OLS in political science is unfortunate given that techniques to estimate dynamic panel data models continue to be developed, particularly in economics.[35] We have left many important issues and methods largely untouched, including random coefficient models, cointegration, unit-root testing, instrumental variables estimation and Bayesian hierarchal models. Although we certainly sympathize with applied researchers looking for a simple, robust and widely accepted approach to estimating dynamic models, a desire to apply a simple solution to a complex problem does not mean it is appropriate to do so. At the very least, researchers should acknowledge the highly unreliability of dynamic results, as demonstrated in the previous section.

[34] It should be noted here that the real culprit is putting an endogenous variable (the LDV) as a regressor, not from incorporating exogenous unit effects.
[35] In a recent survey, Maddala (1998) gives expert assessment of the field, including some guidance on how these methods apply to political research.

4.2 Some general recommendations

Given this state of affairs, how can the profession move forward? It seems clear that more attention needs to be paid to the following six issues:

1) A **better understanding** by applied researchers of the vast literature related to panel data models can only help. Hopefully, our very brief review is a modest step in this direction. There is *a lot* "to do and not to do" with TSCS data.

2) The wide variety of estimators proposed in the literature need to be thoroughly evaluated with **Monte Carlo analysis** on simulated data sets that bear the properties of TSCS data in the discipline. Ideally, this will be a coordinated, discipline-wide enterprise. The analysis of B&K (1996) is a good first step in this direction, but that analysis was limited and only compared the LDV model with the AR1 within a particular data structure, ignoring a variety of other issues such as unit heterogeneity.

3) We have clearly shown the need for **more sensitivity analysis**. Many papers perform sensitivity analysis with respect to how variables are operationalized and which variables are included in the model, but relatively few conduct sensitivity analysis with respect to the basic estimating equations, particularly with respect to dynamics.

4) Sensitivity analysis will usually reveal a variety of estimates that may seem plausible given theoretical conditions. Therefore, **improved model selection techniques** should be explored. Standard goodness-of-fit tests can be useful, but should not be considered definitive. Bayesian statisticians have for some time used formal model selection techniques that may be useful to avoid post-hoc rationalization based on goodness-of-fit measures.

5) A **stronger theoretical foundation** is necessary for model selection. Unfortunately, much theorizing in political science is not formalized enough to yield any guidance on model specification. Regardless of our statistical tools, we can never improve the validity of our results from a scientific perspective without stronger theories.

6) We need to increase the practice of **conservative, scientific inference**. Editors, reviewers, and publication-seeking researchers must all acknowledge that the "best answer" for a particular question may be that there is *no* reliable answer attainable given the data at hand.

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. B&K's work implies that the LDV model is the most straightforward dynamic approach, but we see nothing about the LDV that makes it inherently more simple than other specifications (which, our replications show, yield sharply different results in many cases). 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 to swallow—many data sets simply have too many limitations to use in a reliable fashion (without illegitimately inflating sample size, for instance).

The general frailty of TSCS analyses in the literature is even more disconcerting when we consider that the bias associated with dynamic panel models discussed in Section 2.4 has not been adequately appreciated by almost all of the published work in political science. It is possible that the actual biases in real data (as opposed to the theoretical bias we point out) turn out to be trivial (Judson and Owen (1999) suggests they might be), but so far there is not sufficient evidence for such optimism.

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 nor reference to the voluminous literature on dynamic panel data modeling that existed long before 1995—led hundreds of 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. Since political science is a discipline in which many practitioners are not trained to read the technical statistical literature, they depend on sound and cautious methodological advice. Unfortunately, practitioners have not been well served. Ultimate responsibility, of course, relies on researchers who felt satisfied simply to quote a prominent article for support of their method, rather than deriving support from carefully dealing with the issues presented in the literature. The opportunity cost of this misfortune is that numerous papers (both good ones and ones not so good) could have been much stronger than they were.

The recommendations we give above focus entirely on statistical analysis. But given the type of data being used, it is likely that statistical robustness will remain unattainable in many cases. Although continued statistical analysis and the development of better methods are essential, 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. And of course the need for more rigorous theory with strong empirical predictions should not be understated.

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

Achen, Christopher H. 2000. "Why Lagged Dependent Variables can Suppress the Explanatory Power of Other Independent Variables." Paper presented at the Annual Meetings of the Political Methodology Section of the American Political Science Association, University of California, Los Angeles, 20-22 July.

Ahn, Seung C., and Peter Schmidt. 1995. "Efficient Estimation of Models for Dynamic Panel Data." *Journal of Econometrics* 68(July): 5-27.

Arellano, Manuel. 1989. "A Note on the Anderson-Hsiao Estimator for Panel Data." *Economic Letters* 31: 337-341.

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(April): 277-297.

Beck, Nathanial. 2001. "Time-Series—Cross-Sction 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(September):634-647.

---------- 1996. "Nuisance v. Substance: Specifying and Estimating Time-Series—Cross-Section Models." *Political Analysis* 6:1-36.

---------- 2001. "Throwing out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon. *International Organization* 55(spring): 487-495.

Bun, Maurice J.G., and Jan F. Kiviet. 2001. "The Accuracy of Inference in Small Samples of Dynamic Panel Data Models: Simulation Evidence and Empirical Results." Working Paper.

Bun, Maurice J.G. and Martin A. Carree. 2002. "Bias-Corrected Estimation in Dynamic Panel Data Models." Working Paper.

Carree, Martin A. 2001. "Nearly Unbiased Estimation in Dynamic Panel Data Models." Working Paper.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *The Review of Economic Studies* 47(January): 225-238.

Cox, Gary W., Frances M. Rosenbluth, and Michael F. Thies. 1998. "Mobilization, Social Networks, and Turnout: Evidence from Japan." *World Politics* 50(April): 447-474.

Green, Donald P., Soo Yeon Kim, and David H. Yoon. 2001. "Dirty Pool." *International Organization* 55(spring): 441-468.

Greene, William H. 2000. *Econometric Analysis, 4th Edition*. Upper Saddle River, NJ: Prentice-Hall.

Gujarati, Damodar N. 1995. "Dynamic Econometric Models: Autoregressive and Distributed Lag Models. In *Basic Econometrics, 3rd Edition.* New York: Mcgraw-Hill Inc. pp. 584-634.

Hausman, Jerry A. 1978. Specification Tests in Econometrics. *Econometrica* 46: 1251-1271.

Hausman, Jerry A., and William E. Taylor. 1981. "Panel Data and Unobservable Individual Effects." *Econometrica* 49(November): 1377-1398.

Heckman, James J. 1981. "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Series Data Stochastic Process." In *Structural Analysis of Discrete Data*, eds. Charles F. Manski and Daniel McFadden. Cambridge: The MIT Press. Pp. 179-195.

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(November): 599-621.

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(July): 107-150.

Judson, Ruth A., and Ann L. Owen. 1999. "Estimating Dynamic Panel Data Models: A Guide for Macroeconomists." *Economic Letters* 6: 9-15.

Kittel, Bernhard. 1999. “Sense and Sensitivity in Pooled Analysis of Political Data.” *European Journal of Political Research* 35(March): 225-253.

Kiviet, Jan F. 1995. “On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models.” *Journal of Econometrics* 68(July): 53-78.

Kristensen, Ida P. and Gregory Wawro (2003). “Lagging the Dog? The Robustness of Panel Corrected Standard Errors in the Presence of Serial Correlatoin and Observation Specific Effects.” Unpublished Manuscript, Columbia University.

Moene, Karl Ove, and Michael Wallerstein. 2001. “Inequality, Social Insurance, and Redistribution.” *American Political Science Review* 95(December):859-874.

Nerlove, Marc. 1971. “Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections.” *Econometrica* 39(March): 359-382.

Nickell, Stephen. 1981. “Biases in Dynamic Models with Fixed Effects.” *Econometrica* 49(November): 1417-1426.

Parks, Richard. 1967. “Efficient Estimation of a System of Regression Equations When Disturbances are Are Both Serially and Contemporaneously Correlated.” *Journal of the American Statistical Association* 62:500-509.

Pickering, Jeffrey. “War-Weariness and Cumulative Effects: Victors, Vanquished, and Subsequent Interstate Intervention.” *Journal of Peace Research* 39(3): 313-337.

Poe, Steven C., and C. Neal Tate. 1994. “Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis.” *American Political Science Review* 88(December):853-872.

Reich, Gary M. 1999. “Coordinating Restraint: Democratization, Fiscal Policy and Money Creation in Latin America.” *Political Research Quarterly* 52(December): 729-751

Rueda and Pontusson. 2001. “Wage Inequality and Varieties of Capitalism.” *World Politics* 52(April): 350-383.

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(February): 103-129.

Steenbergen, Marco R., and Bradford S. Jones. 2002. “Modeling Multilevel Data Structures.” *American Journal of Political Science* 46(January):218-237.

Wawro, Gregory. 2002. “Estimating Dynamic Panel Models in Political Science.” *Political Analysis* 10(1): 25-48.

Western, Bruce. 1998. “Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach.” *American Journal of Political Science* 42(October):1233-1259.

Zahariadis, Nikolas. 2001. “Asset Specificity and State Subsidies in Industrialized Countries.” *International Studies Quarterly* 45(December):603-616.

Ziliak, James P. 1997. “Efficient Estimation with Panel Data When Instruments are Predetermined: An Empirical Comparison of Moment-Condition Estimators.” *Journal of Business & Economic Statistics* 15(October): 419-431.

Zorn, Christopher W. 2001. “Generalized Estimating Equation Models for Correlated Data: A Review with Applications.” *American Journal of Political Science* 45(April):470-490.