BEYOND "FIXED VERSUS RANDOM EFFECTS": A FRAMEWORK FOR IMPROVING SUBSTANTIVE AND STATISTICAL ANALYSIS OF PANEL, TIME-SERIES CROSS-SECTIONAL, AND MULTILEVEL DATA

Brandon L. Bartels
Assistant Professor
Department of Political Science
George Washington University
Washington, DC 20052
bartels@gwu.edu

Abstract

Researchers analyzing panel, time-series cross-sectional, and multilevel data often choose between random effects, fixed effects, or complete pooling modeling approaches. While pros and cons exist for each approach, I contend that some core issues continue to be ignored. I propose a modeling framework for analyzing clustered data that solves various substantive and statistical problems. The approach: (1) solves the substantive interpretation problems associated with *cluster confounding*, which occurs when one assumes that within- and between-cluster effects are equal; (2) accounts for cluster-level unobserved heterogeneity; (3) satisfies the controversial statistical assumption that level-1 variables be uncorrelated with the random effects term; (4) allows for the inclusion of level-2 variables; and (5) allows for statistical tests of cluster confounding. I illustrate this approach using three substantive examples: global human rights abuse, oil production for OPEC countries, and Senate voting on Supreme Court nominations. Analyses highlight how the proposed framework enhances substantive interpretations.

A previous version of this paper was prepared for the Faculty Poster Session, Political Methodology Conference, Ann Arbor, MI, July 9-12, 2008. I thank the authors whose work I reexamine in the substantive applications for making their data available online.

Political scientists analyzing clustered data—namely panel, time-series cross-sectional (TSCS), and multilevel (or hierarchical) data—face difficult choices when confronting the model specification and estimation stages of their research. Importantly, clustered data structures possess multiple levels of analysis where lower-level units of analysis are nested within higher-level units of analysis. Clustering induces unobserved heterogeneity across clusters, meaning the conditional cluster means of the dependent variable vary for unobserved reasons. To examine clustered data, political scientists often choose between a "fixed effects" (FE), "random effects" (RE),[1] and "complete pooling" modeling approach. The first two approaches account for unobserved heterogeneity, though in very different ways, while complete pooling ignores unobserved heterogeneity altogether. Moreover, each approach produces different, and, in some cases, ambiguous substantive interpretations of coefficients.

While debates continue within political science about which approach is best for certain situations (e.g., Beck 2001; Beck and Katz 2001, 2007; Green, Kim, and Yoon 2001; Stimson 1985; Wilson and Butler 2007), I argue that some core issues concerning clustered data continue to be both mischaracterized and ignored. In addition to clarifying some misconceptions about extant approaches, I present a unified and simple modeling framework for analyzing clustered data, which should be of general interest to analysts of panel, TSCS, and multilevel data. I call this a "unified" approach because it solves many of the substantive and statistical problems that extant approaches possess. First, the method solves the problem of *cluster confounding*, which occurs when a level-1 variable (a time-varying variable in TSCS and panel data) exhibits distinct

[1] I use "random effects" and "random intercept" interchangeably throughout the paper. The term "random effects" technically implies both a random intercept model and the more general random coefficient model. But most people refer to the former when they use the term.

within-cluster and between-cluster effects, yet one does not distinguish these two types of variation in the variable. Thus, the within- and between-cluster effects are combined, or confounded, together into a single effect (e.g., Skrondal and Rabe-Hesketh 2004; Zorn 2001b). The solution, which entails estimating separate within- and between-cluster effects, allows for more explicit substantive interpretations of effects. Second, estimation of a random intercept model (or more generally, a random coefficient model) allows one to control for unobserved heterogeneity at the cluster level. Third, the solution to cluster confounding satisfies the controversial statistical assumption associated with the RE approach that level-1 independent variables be uncorrelated with the random effects term. Fourth, unlike the FE approach, the proposed method allows for the inclusion of level-2 variables (time-constant variables in TSCS and panel data), thus not limiting the types of hypotheses one can test. And fifth, the method allows for statistical tests of cluster confounding, i.e., whether differences between within- and between-cluster effects are statistically significant.

I empirically illustrate the modeling approach using three substantive examples: (1) global human rights abuse (Poe and Tate 1994; Poe, Tate, and Keith 1999); (2) oil production in OPEC countries (Blaydes 2004, 2006; Goodrich 2006); and (3) Senate voting on Supreme Court nominations (Epstein, Lindstadt, Segal, and Westerland 2006). Reexaminations of these data produce refined interpretations of the some of the core substantive conclusions.

CLUSTERING AND UNOBSERVED HETEROGENEITY

As is well known, clustering induces *unobserved heterogeneity*, which means that the cluster means of the dependent variable will vary across clusters because of unmeasured cluster-level factors. Unobserved heterogeneity is a core concept that should always be addressed in clustered data. For some models, one can include observed variables that will explain part of this variation in the dependent variable across clusters, but there will almost always be residual error variance at the cluster level, just as there is always residual error variance in a plain vanilla OLS model. Figure 1 provides a simple illustration of unobserved heterogeneity in clustered data. Across the horizontal axis are ten clusters, e.g., individuals or countries in panel or TSCS data; schools, countries, or states in multilevel data. The dots represent values of the dependent variable for each unit of analysis within a given cluster. Each cluster contains six observations. For panel and TSCS data, the dots could represent values of *Y* over six time periods. For multilevel data, the dots could represent six individuals per school. The dash within each cluster represents that cluster's mean of the dependent variable. When statisticians and political methodologists speak of "unobserved heterogeneity" in clustered data, they are simply referring to variation in these dashes across clusters.[2] That is, there is something about cluster 7 that makes it on average higher in values of the dependent variable than clusters 1, 2, and 10; but this "something" cannot be completely captured by observed independent variables. Once some observed cluster-specific independent variables are included in a model, we are then interested in variation in the conditional cluster means of the dependent variable.

[Figure 1 about here]

To introduce these issues in equation form, I prefer a generalized multilevel modeling setup. For now, I assume a linear modeling framework.

(1a) $Y_{ij} = \beta_{0j} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + e_{ij}$ [Level-1 Equation]

(1b) $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_{1j} + u_{0j}$ [Level-2 Equation]

[2] More specifically, I am referring here to unobserved heterogeneity *in the response*. Random coefficient models can account for this type of heterogeneity as well as unobserved *causal* heterogeneity, which means that level-1 effects vary over clusters due to unmeasured factors.

Equations 1a and 1b can be rewritten in a reduced-form representation by substituting the level-2 equation into the level-1 equation:

$$(2) \qquad Y_{ij} = \gamma_{00} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \gamma_{01} Z_{1j} + u_{0j} + e_{ij}$$

In this setup, *i* indexes level-1 units and *j* indexes level-2 units. In TSCS and panel data, *i* represents measurement occasions and *j* represents individuals or countries. TSCS and panel data modelers are used to communicating the number of cross-sectional units (*N*) and time points (*T*). In the multilevel representation above, cross-sectional units are level-2 units and *T* represents the cluster sizes for each cluster (the number of measurement occasions per cluster). Thus, if *N*=30 and *T*=40, we have 1,200 measurement occasions (level-1 units) nested within 30 individuals or countries (level-2 units). Two variables, X_{1ij} and X_{2ij}, are included at level 1. For panel and TSCS data, these are time-varying variables. For multilevel data, with individuals nested within higher-level units, these would be individual-level variables. Z_{1j} is a level-2 variable, which is a time-constant (or country/individual-specific) variable in panel and TSCS data and a contextual variable in multilevel data. e_{ij} represents the level-1 error, a random term assumed to be normally distributed with mean zero and an estimable variance.

The inclusion of β_{0j} means that the intercept is allowed to vary somehow across level-2 units. u_{0j} represents unobserved heterogeneity across clusters, and as I discuss in more detail below, there are alternative ways to treat u_{0j}. Referring to back to Figure 1, the inclusion of β_{0j} and u_{0j} allows the conditional means of the dependent variable to vary across level-2 units for unobserved reasons. Note how the level-2 equation allows for the varying intercept to be explained by observed (Z_{1j}) and unobserved heterogeneity (u_{0j}). Failure to account for unobserved heterogeneity (i.e., completely pooling the data) forces the conditional cluster means of the dependent variable to be equal, which is a restrictive assumption indeed, though one we can test for. If violated, forcing this assumption will lead to biased parameters estimates (e.g., Gelman and Hill 2007; Hsiao 2003; Raudenbush and Bryk 2002; Skrondal and Rabe-Hesketh 2004). For panel and TSCS data, dynamics are also a concern, and other work discusses this issue in greater depth (Beck and Katz 1996; Hsiao 2003; Heckman 1981; Wilson and Butler 2007). Throughout this paper, I adopt a standard practice of using a lagged dependent variable to account for dynamics.

$$(3) \qquad Y_{ij} = \gamma_{00} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 Y_{ij(t-1)} + \gamma_{01} Z_{1j} + u_{0j} + e_{ij}$$

MODELING APPROACHES FOR HANDLING UNOBSERVED HETEROGENEITY

How to model unobserved heterogeneity (u_{0j}) in clustered data constitutes a core debate in the statistical literature generally (e.g., Baltagi 2005; Hsiao 2003; Wooldridge 2002) and in political science applications (e.g., Beck 2001; Green et al. 2001; Stimson 1985; Wilson and Butler 2007; Zorn 2001a). Three general avenues are available for treating u_{0j}. The first is a *complete pooling* approach, which assumes that u_{0j}=0 and thus ignores unobserved heterogeneity. Note that a commonly-used modeling approach for TSCS data—the original Beck and Katz (BK) (1995) recommendation of using "panel-corrected standard errors" (PCSEs)—is a complete pooling approach that does not account for unobserved heterogeneity. The authors' PCSEs do, of course, make corrections for the standard errors, but the OLS coefficients that the authors recommend are completely pooled estimates. The major payoff of this approach is its simplicity, and numerous practitioners have implemented this procedure (see Wilson and Butler 2007 for an extensive review). A disadvantage of complete pooling is that ignoring unobserved heterogeneity can induce omitted variable bias (e.g., Hsiao 2003; Skrondal and Rabe-Hesketh 2004). Moreover, as I will discuss in more detail below, interpretation of results is unclear because the coefficients assume that the within- and between-cluster effects are equal. Thus, one cannot be completely confident over which level of analysis (e.g., longitudinal versus cross-sectional; individual versus aggregate) the relationship actually occurs.

Second, a fixed effects approach allows each level-2 unit to possess its own intercept, meaning u_{0j} is treated as fixed. The FE approach is a "no pooling" approach. Since the cluster dummies absorb all of the between-cluster variation in the data, the effects of X_{1ij} and X_{2ij} are solely within-cluster effects and the effect of Z_{1j} cannot be estimated. For TSCS data, a now standard modeling practice is to use an FE model with panel-corrected standard errors and a lagged dependent variable to account for dynamics (Beck and Katz 1996; Beck 2001; Wilson and Butler 2007), though there is not an ironclad consensus about this strategy among practitioners (see, e.g., Blaydes 2006; Goodrich 2006).

One of the concerns practitioners raise about the FE model is that it eats up too many degrees of freedom, resulting in shaky estimates (e.g., Beck 2001; Beck and Katz 2001). This is somewhat of a misconception. Since all between-cluster variation in the data is absorbed by the cluster-specific dummies, the effects of independent variables are solely within-cluster effects, which has implications for how one interprets coefficients. For TSCS data, such effects are interpreted as: for a given country, as X varies *across time* by one unit, Y increases or decreases by β units. The fact that cluster-specific independent variables (like Z_{1j} in equation 1b) cannot be included in the FE model is seen as a major disadvantage of the FE approach since it eliminates the ability to test between-cluster hypotheses. Another disadvantage is that one cannot retrieve "good" estimates of sluggish, or slowly-changing, variables in the FE model (Beck 2001; Plumper and Troeger 2007). Though again, this should not be surprising, since the FE model produces solely within-cluster effects. For variables in panel or TSCS data that do not vary much over time, we should expect coefficients to be inefficient given the lack of within-cluster information in the data. For sluggish variables, the issue is not the FE model, per se, but instead with the nature of the data. Importantly, one can test for the adequacy of the FE specification by performing a joint *F*-test of the cluster dummies.

Third, a random effects, or random intercept, approach treats u_{0j} as distributed normally with mean zero and an estimable variance. This approach decomposes the total error into a level-1 component (e_{ij}) and a level-2 component (u_{0j}). The RE model is a "partial pooling" approach, with the effects of X_{1ij} and X_{2ij} a weighted average of the within and between-cluster variation in the data (e.g., Gelman and Hill 2007; Hsiao 2003; Skrondal and Rabe-Hesketh 2004). The RE approach, and the more generalized random coefficient model, is widely used in analyses of panel data (with large *N* relative to *T*) and multilevel data (e.g., Bowler, Donovan, and Hanneman 2003; Martin 2001; Steenbergen and Jones 2002).[3]

A major complaint lodged against the RE model relates to the restrictive assumption that level-1 independent variables be uncorrelated with the random effects term: $Cov(X_{ij}, u_{0j})=0$. Since a level-1 variable varies both within and between clusters, many argue that this an unrealistic assumption to satisfy, since unobserved heterogeneity will almost always be correlated with the independent variables. This controversial assumption often makes the FE model, which does not incorporate this assumption, a superior choice over the RE model (e.g., Beck 2001; Kristensen and Wawro 2003; Wilson and Butler 2007). Analysts often rely on the Hausman (1978) test to assess the adequacy of this controversial assumption. I will have more to

[3] GEE models (Zeger and Liang 1986; Zorn 2001a), which share some basic similarities to the RE approach, are becoming more commonly used in political science analyses of clustered data. Also called population-averaged models, GEE estimates are marginal with respect to unobserved heterogeneity, while RE estimates are conditional with respect to unobserved heterogeneity.

say about this assumption and the Hausman test later on. Moreover, for TSCS data, some argue that RE is inappropriate because it treats unobserved heterogeneity across countries as random, yet for a population of countries, an FE approach would be superior (Beck 2001; Kristensen and Wawro 2003). This is a major misconception. Unobserved heterogeneity represents unmeasured *differences* between countries in the dependent variable, and so the RE approach simply separates random error into a within-cluster (e_{ij}) and between-cluster (u_{0j}) component. The latter represents random error across, e.g., countries, just as a simple OLS model would contain a random error term that captures unobserved differences across countries in a cross-sectional analysis. A disadvantage of the RE approach—and one shared with the complete pooling approach—relates to the interpretation of coefficients. Though the coefficients from an RE model are now partially pooled, as opposed to completely pooled, the estimates still assume that the within- and between-cluster effects are equal, thus making substantive interpretations imprecise. One major advantage of the RE approach over FE is that one can include level-2 variables (e.g., time-constant variables in TSCS and panel data), which allows one to test the effects of between-cluster variables.

A UNIFIED MODELING APPROACH FOR CLUSTERED DATA

In this section, I elaborate on a simple yet powerful methodology capable of solving many of the substantive and statistical problems common to extant approaches and, at the same time, maintaining many of the positive aspects of these approaches. The approach: (1) solves the substantive interpretation problems associated with *cluster confounding*, which occurs when one assumes that the within- and between-cluster effects are equal; (2) accounts for unobserved heterogeneity via the use of a random intercept model, which incorporates a random error at the cluster level; (3) satisfies the controversial statistical assumption that the level-1 variables be uncorrelated with the random effects term; (4) allows for the inclusion of level-2 variables, something the FE approach cannot accommodate; and (5) allows for statistical tests of cluster confounding, which bear resemblance to the Hausman test.

Cluster Confounding. To motivate the issue of cluster confounding, I draw upon work in statistics by Skrondal and Rabe-Hesketh (2004, 50-53; see also Rabe-Hesketh and Skrondal 2005) and in political science by Zorn (2001b). In panel, TSCS, and multilevel data, there are multiple sources of variation in the data, which has implications for how we understand the effects of independent variables. It is worth remembering that variables may be measured so that (1) they vary both within and between clusters (e.g., time-varying variables in panel and TSCS data; level-1 variables in multilevel data) or (2) they vary only between clusters and not within clusters (e.g., time-constant variables in panel and TSCS data; level-2 variables in multilevel data). Relationships between independent variables and the dependent variable will vary over different units of analysis depending on which level they are measured at. An important issue that has gotten lost in the debate over modeling approaches is the notion that a level-1 variable may exhibit quite distinct within- and between-cluster effects, as highlighted by Zorn (2001b) in the context of discrete-time duration modeling. For example, in TSCS data, what if X exhibited a null within-cluster, or longitudinal, effect but a positive between-cluster effect? We would conclude that, for a given country, increases in X over time do not affect Y. But across countries, as average levels of X increase, average levels of Y increase as well. Recall that the FE model would only recover the within-country effect. Importantly, the complete pooling and RE models would assume that the within- and between-country effects are equal. That is, we would have one coefficient, and we would assume that, for a given country, a one-unit change in X across time has the same impact on Y as a one-unit change in the average of X between countries.

The example above is one of *cluster confounding*, which occurs when a level-1 variable exhibits distinct within-cluster and between-cluster effects, yet one only includes the original level-1 variable in the model without distinguishing these two types of variation in the variable. As a result of not making this distinction, the within- and between-cluster effects are combined, or confounded, into a single effect representing an average of the within- and between-cluster effects. If the within- and between-cluster effects of a level-1 variable are the same, which is something we can test for, then cluster confounding is not a problem. But if they are not equal, the uncorrected results cannot distinguish whether the effects are within- or between-cluster effects. Cluster confounding has significant implications for how one interprets the effects of independent variables in clustered data, and therefore, detecting and correcting for it is crucial for understanding the precise nature of relationships and for testing hypotheses.

Figure 2 illustrates the importance of cluster confounding by presenting different types of scenarios of within- versus between-cluster effects. For each plot, three clusters (e.g., countries, individuals, states, schools) are presented. The solid lines represent within-cluster effects and the dashed lines represent between-cluster effects. In Figure 2A, no cluster confounding exists; the within- and between-cluster slopes are equal. In the remaining three plots, significant cluster confounding occurs. Figure 2B presents a scenario where there is a positive within-cluster effect, but a negative between-cluster effect. For TSCS data, this would mean that, for a given country, increases in *X* produce increases in *Y*, but between countries, as average levels of *X* increase, the average of *Y* does not change. Figure 2C illustrates drastic cluster confounding, where the within-cluster effect is positive, but the between-cluster effect is negative. And Figure 2D represents a scenario where the within-cluster effect is null, while the between-cluster effect is negative. For instance, in multilevel data, we might have a null individual effect but a negative aggregate effect. One can imagine additional cluster confounding scenarios as well. On the whole, Figure 2 highlights the dire consequences of not accounting for cluster confounding in empirical analysis. One runs the risk of making incorrect substantive interpretations and rendering incorrect verdicts on hypotheses.

[Figure 2 about here]

Solving the problem of cluster confounding first involves calculating within- and between-cluster transformations of a level-1 variable, X_{ij} (e.g., Skrondal and Rabe-Hesketh 2004; Zorn 2001b). One first calculates the cluster-specific mean of X_{ij}, which we will call $\overline{X}_j$. This is the between-cluster operationalization of X_{ij}. Then, the within-cluster operationalization of X_{ij} is calculated as: $X_{ij}^W = X_{ij} - \overline{X}_j$. Since we have completely separated the within from the between-cluster variation in X_{ij}, note that $\overline{X}_j$ and X_{ij}^W are completely uncorrelated. As I discuss in more detail in the substantive applications, X_{ij}^W represents deviations in units of measurement from the cluster mean. I have created a Stata program to generate these within- and between-cluster transformations. Details are in Online Appendix A.

Random Intercept Model. The next step involves specifying a random intercept model and including the within- and between-cluster transformations of the *X*'s in the model. Importantly, this modeling approach solves the problem of cluster confounding while accounting for cluster-level unobserved heterogeneity. I use the reduced-form representation of the model from equation 2 to demonstrate the approach:

$$(4) \qquad Y_{ij} = \gamma_{00} + \beta_1 X_{1ij}^W + \beta_2 X_{2ij}^W + \gamma_{01} Z_{1j} + \gamma_{02} \overline{X}_{1j} + \gamma_{03} \overline{X}_{2j} + u_{0j} + e_{ij}$$

β_1 and β_2 now represent *within-cluster effects* of X_1 and X_2, respectively. These would be purely individual effects in typical multilevel data and purely longitudinal effects for TSCS and panel data. γ_{02} and γ_{03} now represent *between-cluster effects* of X_1 and X_2, respectively. These would be aggregate effects in multilevel data and cross-sectional effects in panel and TSCS data. In the substantive applications, I discuss interpretations of these effects in more detail. Since this is a random-intercept model, the total error is partitioned into a within-cluster (e_{ij}) and between-cluster component (u_{0j}). Both are assumed to be normally distributed with means equaling zero and estimable variances.

Advantages. An extremely important feature of this model is that it satisfies the controversial assumption, $Cov(X_{ij}, u_{0j})$=0. The within-cluster transformations of X_1 and X_2 are now completely uncorrelated with the between-cluster random effect, u_{0j}, thus escaping the bias that can occur when violating this assumption. Of course, we still assume that all level-2 variables are uncorrelated with u_{0j} (e.g., $Cov(\overline{X}_{1j}, u_{0j}) = 0$), but then again, we make similar assumptions in a simple OLS regression that the independent variables be uncorrelated with the error term. Another important feature of this model is that, unlike the FE model, one can include level-2 variables, like Z_{1j}, in the model. Thus, the major advantage of this modeling approach over the FE model is that one can still estimate within-cluster effects of variables, but in addition, one can simultaneously estimate between-cluster effects and the effects of additional level-2 variables. Unlike the FE model, the proposed approach does not limit the types of hypotheses one can test.

Statistical Tests for Cluster Confounding. Another important feature of this modeling approach is that it allows for statistical tests of whether cluster confounding poses a significant problem, that is, whether the differences between the within- and between-cluster effects are statistically significant. To perform these tests, one estimates the same underlying model as in equation 4 but with different operationalizations of the *X*'s:

$$(5) \qquad Y_{ij} = \gamma_{00} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \gamma_{01} Z_{1j} + \gamma_{02} \overline{X}_{1j} + \gamma_{03} \overline{X}_{2j} + u_{0j} + e_{ij}$$

Instead of including the within-cluster operationalizations of the *X*'s (i.e., X^W_{1ij} and X^W_{2ij}) as was done in equation 4, one includes the originally-coded X_{1ij} and X_{2ij}.[4] For this specification, β_1 and β_2 in equation 5 will be identical to β_1 and β_2 in equation 4; they still represent *within-cluster effects* of X_1 and X_2, respectively. However, including the originally-coded *X*'s instead of the within-cluster transformations changes the meaning of γ_{02} and γ_{03}. In equation 5, γ_{02} and γ_{03} now represent the *differences* between the within-cluster and between-cluster effects of X_1 and X_2, respectively (see Skrondal and Rabe-Hesketh 2004, 53). These γ coefficients, in conjunction with their standard errors, allow one to test for the existence of cluster confounding, that is, whether the differences between the within- and between-cluster effects are statistically significant.

Note the resemblance of this testing procedure to the Hausman (1978) test, which tests for differences between coefficients from an FE model an RE model. The Hausman test essentially assesses the adequacy of the RE model's assumption that the within- and between-cluster effects are equal. If they are equal, then cluster confounding is not a problem, and therefore the RE coefficients will not differ systematically from the FE coefficients. Many practitioners also conclude that significant differences between FE and RE estimates means that the RE estimates are inconsistent due to the violation of the controversial assumption that $Cov(X_{ij}, u_{0j})=0$. But the estimation of distinct within- and between-cluster effects for *X* removes this bias in the RE model (see, e.g., Skrondal and Rabe-Hesketh 2004, 52-53, 269).

A Note on Dynamics. To account for dynamics in TSCS and panel data, one should add

[4] Note that this specification is the approach suggested by Bafumi and Gelman (2006). However, the authors are not explicit about the interpretations of the coefficients.

the within-cluster operationalization of the lagged dependent variable, $Y^{W}_{ij(t-1)}$. This represents how, for a given country, past values of the dependent variable influence current values. It would not be substantively meaningful, however, to include the lagged cluster mean of Y_{ij} (i.e., $\overline{Y}_{j(t-1)}$).

Extensions

A logical extension to the random intercept specification discussed above is to specify a random coefficient model (RCM) (see Beck and Katz 2007; Raudenbush and Bryk 2002; Skrondal and Rabe-Hesketh 2004; Steenbergen and Jones 2002; Western 1998). In addition to allowing the intercept to vary across clusters, the RCM allows level-1 coefficients to vary across clusters. Substantively, this accounts for cluster-level heterogeneity in the *effects* of level-1 variables (i.e., causal heterogeneity). If one is interested in how contextual variables shape the magnitude of level-1 effects, one can include cross-level interactions. Below is an illustration:

(6a) $$Y_{ij} = \beta_{0j} + \beta_{1j} X^{W}_{1ij} + \beta_{2j} X^{W}_{2ij} + e_{ij}$$ [Level-1 equation]

(6b) $$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_{1j} + \gamma_{02} \overline{X}_{1j} + \gamma_{03} \overline{X}_{2j} + u_{0j}$$ [Level-2 equations]

(6c) $$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_{1j} + u_{1j}$$

(6d) $$\beta_{2j} = \gamma_{20} + \gamma_{21} Z_{1j} + u_{2j}$$

The within-cluster effects of X_1 and X_2 (i.e., β_{1j} and β_{2j}) are allowed to vary across clusters. Z_{1j}, a level-2 variable, is specified to moderate the impact of the within-cluster effects of X_1 and X_2. Unobserved heterogeneity in the effects of β_1 and β_2 is represented by u_{1j} and u_{2j}, respectively.

Another specification that may be of substantive importance is to model how between-cluster variation in X moderates the within-cluster impact of X. In the multilevel context, Gelman, Shor, Bafumi, and Park (2006) have shown how the individual-level effect of income on vote choice depends on aggregate levels of income across states. That is, within poorer states, poor individuals are significantly more likely to vote Democratic than rich individuals. But within richer states, income essentially has a null individual-level effect. In short, aggregate income across states moderates the individual-level effect of income. A generalized version of such a model can be specified as:

(7a) $Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij}^{W} + \beta_{2j} X_{2ij}^{W} + e_{ij}$ [Level-1 equation]

(7b) $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_{1j} + \gamma_{02} \overline{X}_{1j} + \gamma_{03} \overline{X}_{2j} + u_{0j}$ [Level-2 equations]

(7c) $\beta_{1j} = \gamma_{10} + \gamma_{11} \overline{X}_{1j} + u_{1j}$

(7d) $\beta_{2j} = \gamma_{20} + \gamma_{21} \overline{X}_{2j} + u_{2j}$

In this model, the between-cluster *X*'s moderate their respective within-cluster effects of the *X*'s. The RCM offers additional opportunities for testing substantively important phenomena.

Estimation

The linear random intercept model can be estimated via feasible generalized least squares (FGLS), maximum likelihood (ML), or Bayesian simulation via Markov Chain Monte Carlo (MCMC); technical details of these procedures are discussed extensively elsewhere (Beck and Katz 2007; Gelman and Hill 2007; Hsiao 2003; Skrondal and Rabe-Hesketh 2004; Western 1998). Each procedure should yield similar statistical inferences (assuming one employs diffuse priors in the MCMC approach). Beck and Katz (2007) show that FGLS has poor finite-sample properties for the RCM, so practitioners should proceed with caution when using this approach.

For nonlinear models (with binary, ordinal, count and other non-continuous outcomes), the two "standards" for estimation are ML and MCMC (e.g., Rodriguez and Goldman 2001). These methods have been shown to be significant improvements over penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) procedures implemented in the software *HLM* (see Rodriguez and Goldman 1995, 2001). For ML, maximizing the likelihood entails acquiring the unconditional distribution of the outcome by integrating out the random effect(s). This can be done using numerical integration via quadrature-based methods (Skrondal and Rabe-Hesketh 2004) or simulated maximum likelihood (Train 2003). Skrondal and Rabe-Hesketh have found that adaptive quadrature produces more accurate results compared to standard quadrature. For RCMs, as the number of random effects increases, ML becomes computationally inefficient, and analysts should consider using MCMC instead. Estimation via ML is available in both Stata (using the "xt" commands) and R (using the "lme" or "nlme" packages). Additional details about model estimation and software are included in Online Appendix A.

Standard Errors. In TSCS analysis, standard errors have received a great deal of attention. Beck and Katz's (1995) panel-corrected standard errors (PCSEs) adjust OLS standard errors for panel heteroskedasticity (due to clustering) and contemporaneous error correlation. Since the proposed framework outlined above accounts for cluster-level heterogeneity and separates within- from between-cluster variation in level-1 variables, threats to the accuracy of standard errors should be minimal. One can always test for various forms of heteroskedasticity that may exist even after modeling heterogeneity, and robust standard errors could be used to correct for any heteroskedasticity that may exist. Online Appendix B provides a further discussion of this issue and an empirical comparison highlighting how standard errors from the proposed framework produce highly similar inferences to those from alternative models that explicitly correct for standard errors.

EMPIRICAL ANALYSIS: THREE SUBSTANTIVE APPLICATIONS

To illustrate the proposed methodology, I present three substantive applications. I use two TSCS applications, both of which involve estimation of linear models. The first is a reexamination of global human rights abuse (Poe and Tate 1994; Poe, Tate, and Keith 1999). The data possess a large *N* relative to *T*, therefore bearing some resemblance to panel data.[5] The

[5] Beck (2001) argues that a key distinction between panel and TSCS data is that the units in panel data (individuals) are sampled from a larger population, while the units in TSCS data

second application is a reexamination of the "rewarding impatience" hypothesis regarding oil production in OPEC countries (Blaydes 2004, 2006; Goodrich 2006). These data contain a small *N* relative to *T*. The third application is a multilevel analysis with a binary dependent variable, where I reexamine Epstein, Lindstadt, Segal, and Westerland's (2006) analysis of Senate voting on Supreme Court nominations. For all three applications, I discuss and present some graphical post-estimation strategies which greatly illuminate substantive interpretations of the results.

Global Human Rights Abuse, 1977-1993

Poe and Tate (1994) provide an important and influential examination of global human rights abuse. Their study is rich with normative, theoretical, and empirical implications. In that analysis, the authors examine 153 countries from 1981 to 1987. In Poe et al. (1999), the authors update and backdate their data over time, add some countries to the dataset, and present new models that refine some of the substantive conclusions from the earlier work. These data include 164 countries covering the years 1977 to 1993. For the dependent variable, the authors rely on "political terror scales," where a country is categorized on a scale of 1 to 5 based on the "occurrence of political imprisonment, execution, disappearances, and torture" (Poe et al. 1999, 297). Countries are categorized by coding the yearly reports from both Amnesty International (AI) and the State Department (SD). Thus, there are two dependent variables, both ranging from 1 to 5, where higher values represent higher levels of personal integrity rights abuse. Results from models using both the AI and SD dependent variables yield similar results. In my reexamination, I analyze the AI dependent variable only.

represent a population of countries. Another major difference is that in TSCS data, *N* and *T* are usually not drastically different, while in panel data, *N* is very large relative to *T*. It is for this latter reason that I say this first example bears some resemblance to panel data.

The authors include ten time-varying variables (i.e., level-1 variables). These include a *lagged dependent variable* (to account for dynamics), *democracy* (7-point Freedom House political rights scale; higher values indicate higher levels off democracy),[6] *population size* (logged population), *population change* (percent change from the previous year), *economic standing* (per capita GNP), *economic growth* (percent change in GNP from the previous year), *leftist government* (dummy variable), *military control* (dummy variable), *international war* (dummy variable), and *civil war* (dummy variable). The authors include one time-constant variable (i.e., level-2 variable): *British cultural influence* (dummy variable). For more details on measurement, see Poe et al. (1999, 296). Using a completely pooled modeling approach (OLS with PCSEs), Poe et al. (1999) find that democracy, economic standing, and British cultural significantly decrease levels of rights abuse, while population size, military control, international war, and civil war significantly increase rights abuse.

I reexamine these results using the modeling approach advocated above. The results are presented in Table 1. The total number of observations is 2,471; N=164, and T≈15.[7] The left side of Table 1 reports a replication of Poe et al.'s results. The right side of the table reports results from a linear random intercept model, estimated via ML, which includes within-country effects, between-country effects, and the absolute value of the difference between the within- and between-country effects (which tests for cluster confounding). Regarding model fit, a likelihood ratio test strongly supports the specification of the random intercept model over a completely pooled approach; significant unobserved heterogeneity (u_{0j}) exists at the country level. The

[6] The authors also used the Polity III democracy scale, and results were very similar to those using the Freedom House measure.

[7] The data are unbalanced, so T varies across countries.

estimate of ρ suggests that 46% of the error variance is accounted for by the country-level error.

[Table 1 about here]

Moving to the results, note first that the within-country lag of the dependent variable exhibits a statistically significant effect, meaning that, for a given country, as past values of rights abuse increase, current values increase as well. Poe et al. find that democracy significantly decreases rights abuse. When decomposing variation in democracy, results reveal that democracy exhibits about equal within-country (-0.10) and between-country (-0.11) effects on rights abuse. And the test of cluster confounding suggests that the difference between these two effects is statistically insignificant. Thus, the effect of democracy can be viewed as a pooled estimate, with the within- and between-country effects being equal, as assumed by Poe et al. Similar results exist for population size. The within- and between-country effects are roughly equal, and the difference between them is statistically insignificant. For a given country, as population size increases over time, rights abuse significantly increases. And, countries with greater populations on average have higher average levels of rights abuse than countries with lower average populations. Poe et al.'s pooled analysis shows that population change has an insignificant effect on rights abuse levels. However, results from the random intercept model reveal that significant cluster confounding occurs for this variable. While population change does not exhibit a significant within-country effect, it does have a statistically significant between-country effect. This means that countries that have undergone greater average levels of population change have significantly greater levels of rights abuse. Economic standing also exhibits significant cluster confounding. Poe et al. report a positive and significant pooled effect of this variable, but my results reveal that per capita GNP exhibits a statistically insignificant within-country effect and a statistically significant between-country effect. Thus, one cannot conclude that as a particular country's per capita GNP increases across time, rights abuse significantly decreases. What we can conclude is that countries with generally higher levels of per capita GNP have generally lower levels of rights abuse. Percent economic change exhibits neither a significant within- nor between-country effect on rights abuse.

Poe et al. report that leftist governments have significantly lower levels of rights abuse compared to non-leftist government. Does this effect occur for a given country across time (as a particular country moves in and out of being a leftist government) or between countries for ones that have been leftist more frequently? The results from the random intercept model support the latter. Since leftist government is a dummy variable, the between-country operationalization is the proportion of the time countries have leftist governments over this time span. Thus, as this proportion increases, rights abuse significantly decreases. The within-country effect is statistically insignificant. Results reveal marginally significant levels of cluster confounding (p=0.10), but given the within effect is insignificant and the between effect is significant, estimating both effects for this variable makes substantive interpretation much more precise. The results for military control resemble those for leftist government. Poe et al. report that military control significantly increases rights abuse. Results from the random intercept model reveal that this pooled effect shows significant cluster confounding. For a given country, a change in the state of military control across time exhibits a null impact on rights abuse. However, as the proportion of years in which a country is under military control increases across countries, rights abuse significantly increases.

The effects of the two war variables show significant cluster confounding. International war exhibits a statistically insignificant within-country effect, meaning that, for a given country, being at war at a given time point does not significantly increase rights abuse compared to not being at war during another time point. However, countries that have been at war more frequently have significantly higher rights abuse levels compared to countries that have been at war infrequently. Civil war exhibits statistically significant within- and between-country effects, though the between-country effect is significantly greater. Finally, countries with British cultural influence (a between-country variable) experience significantly lower rights abuse levels than countries without such influence.

On the whole, the results clarify and refine some of the core conclusions made by Poe et al. It is worth noting that the between-country effects are consistently stronger than the within-country effects, which is sensible given that there is more between-country information in the data (i.e., 164 countries) than within-country information (about 15 years). To illuminate interpretations of these effects, Figure 3 presents graphical depictions of within-country (left column of Figure 3) and between-country effects (right column) for four variables of interest. The graphs depict predicted values ($\hat{Y}$) of the dependent variable while allowing the variable of interest to vary and holding the remaining variables constant at their mean values.[8]

For the between-country effects, the *X*-axis simply represents variation in the country specific means of the particular variable ($\overline{X}_j$). For the within-country operationalizations, where $X_{ij}^{W} = X_{ij} - \overline{X}_j$, units of measurement are now *deviations from the cluster mean*. "0" represents the country mean of the variable for each country. "-1" would represent one unit below the country mean, and "2" would represent 2 units above the country mean. This has implications for how one plots the within-country effects of a variable. In essence, each country will occupy a different range of the within-country measurement space depending on a given country's mean for that variable. To illustrate this issue, consider the democracy variable, which

[8] The cluster-level random effects term, u_{0j}, is set to 0, its expected value.

ranges from 1 to 7. If a particular country's mean of democracy over the time span is 4.6, then that country's within-country operationalization of democracy will range from -3.6 (i.e., 1 – 4.6) to 2.4 (i.e., 7 – 4.6). The within-country slopes for different between-country values will be parallel (as is seen in Figure 3), but changing values of the between-country value will shift the intercept up or down (also illustrated in Figure 3). In Figure 3, I plot within-country effects when the between-country variable is set at low, medium, and high values.[9]

[Figure 3 about here]

Plots A and B in Figure 3 show how the within- and between-country effects of democracy are roughly equal. Both exert rather strong effects, and they highlight how: (1) for a given country, increasing levels of democracy across time significantly reduce rights abuse, and (2) countries that are generally more democratic have significantly lower levels of rights abuse compared to countries that are generally undemocratic. Plots C and D illuminate the null within-country effect and the quite potent between-country effect of economic standing. This distinction has important substantive implications for how we understand the causes of rights abuse. Readers of Poe et al.'s findings may jump to the conclusion that as a given state experiences fluctuations in its economic standing across time, rights abuse will fluctuate as well. But, based on these results, this is an incorrect inference to make. Increases in economic standing for a

[9] To determine what was deemed low, medium, and high values, I had to make judgment calls based on the distributions for each between-cluster variable. *Democracy*: low=2, medium=4, high=6; *economic standing*: low=10th percentile, medium=mean, high=90th percentile; *military control*: low=20th percentile, medium=median, high=80th percentile; *international war*: low=25th percentile, high=75th percentile. The country means of international war have a lopsided distribution, so high and low values were deemed the most appropriate to plot.

given country fail to dampen abuse. Instead, economic standing exhibits only an aggregate, cross-sectional effect, such that countries with generally higher per capita GNP have significantly lower rights abuse levels compared to countries with generally lower per capita GNP. The remaining plots illustrate analogous effects, suggesting that military control and international war exhibit significant between-country effects and insignificant within-country effects. Some might claim that if only a given country would avoid military control, rights abuse would decrease. And, when a given country is at war with another country, rights abuse is higher compared to when that country is not at war. Both are incorrect inferences based on the results. We can only conclude aggregate, between-country effects for these variables. That is, countries that are under military control and in international wars a greater proportion of the time have significantly higher levels of rights abuse. In sum, making these distinctions between within- and between-country effects has important theoretical and empirical implications for our understanding of global rights abuse.

Oil Production in OPEC Countries, 1960-1995

Blaydes (2004) presents empirical evidence in support of her "rewarding impatience" hypothesis, which, derived from a formal bargaining model, posits that impatient countries with shorter time horizons attain significantly greater oil production outputs than patient countries with longer time horizons. Analyzing OPEC countries' oil production levels from 1960-1995 and employing a pooled modeling approach (OLS with PCSEs), Blaydes finds that increases in the amount of per capita oil reserves (the key variable of interest) are associated with significantly lower levels of oil production. She also finds a quadratic effect for per capita reserves, suggesting there is a threshold whereby this "rewarding impatience" effect kicks in. In a response to Blaydes, Goodrich (2006) takes issue with Blaydes's pooled modeling approach and suggests that a fixed-effects approach is superior. Goodrich finds that the within-country effect of the key variable, per capita reserves, is statistically insignificant. But in a separate between-country analysis, Goodrich finds that per capita reserves does indeed exhibit a significant effect. Blaydes (2006) responds to Goodrich with some of the usual criticisms of the FE approach—it is inefficient and that per capita reserves is a "sluggish" variable so the fixed-effect estimate of that variable is inaccurate and inefficient. Recall that the problem related to sluggish variables is not with the FE model, per se, but with the data. If a variable does not greatly vary, one will never retrieve a "good" estimate of that variable unless one collects different and better data. Blaydes estimates a random intercept model, as well as running Plumper and Troeger's fixed-effects vector decomposition (fevd) model for sluggish variables. She contends that the results support her original arguments pertaining to the "rewarding impatience" hypothesis.

Table 2 presents a reexamination of these results. The data consist of 11 OPEC countries over 35 years. The dependent variable is the natural log of annual crude oil production. The independent variables are: *natural log of proven oil reserves*, *natural log of per capita oil reserves* (and a squared term of this variable to test the quadratic effect), a *lagged dependent variable* to account for dynamics, and a *conflict dummy variable* to control for events such as the Iranian revolution, the Iran-Iraq War, the Persian Gulf War, and sanctions on Iraq. All variables are time-varying covariates. The first two models present replications of Blaydes's (2006) Model 3. The first is an OLS model with PCSEs, and the second is a random intercept model.[10] The third model is a random intercept model (estimated via ML) implementing the procedures I have advocated in this paper. Both Blaydes and Goodrich introduce alternative specifications with

[10] I could not produce an exact replication of Blaydes's (2006) Model 3 for both the OLS and random intercept models. The results are similar and produce the same substantive implications.

some additional independent variables, but results from these models produce substantively similar results for the key variables as Model 3.

[Table 2 about here]

In terms of model fit, a likelihood ratio test supports the specification of the random intercept model over a completely pooled approach. There is significant unobserved heterogeneity (u_{0j}) at the country level. Also, the estimate of ρ indicates that 17% of the total error variance is accounted for by the country-level error. Results reveal significant cluster confounding for all variables, suggesting severe discrepancies between the within- and between-country effects of variables. Regarding the key variable of interest, ln(per capita reserves), recall that Blaydes found a negative and significant effect, as well as a negative significant effect for the squared term. This suggests an upside-down U-shaped effect, where there is some threshold at which oil production peaks as a function of per capita reserves. After that threshold, increases in ln(per capita reserves) produce a decrease in oil production. Of course, Blaydes's models assume that the within- and between-country effects of these variables are equal. The random intercept model shows that the within-country effect of both per capita reserves and its squared term are statistically insignificant. In fact, the results suggest a linear, positive within-country effect (as will be seen more clearly in Figure 4), which is contrary to what Blaydes predicted. Thus, for a given OPEC country, increases in ln(per capita reserves) over time produce a positive but statistically insignificant effect on oil production. Turning to the between-country effect of per capita reserves, the results report negative and statistically significant coefficients for both ln(per capita reserves) and its squared term. This means that countries with generally higher levels of per capita reserves have lower oil production than countries with generally lower levels of per capita reserves.

Figure 4 illustrates these findings. Using the same procedures as discussed for Figure 3 (in the human rights abuse example), Figure 4 presents both the within- and between-country effect of ln(per capital reserves) on ln(crude oil production). Note the very small positive within-country effect of per capita reserves on oil production. The quadratic between-country effect of per capita reserves is displayed in plot B. For very low levels of ln(per capita reserves), there is a slightly positive effect. But after the threshold, as average reserves increase between countries, average oil production decreases. On the whole, the "rewarding impatience" hypothesis occurs at the aggregate, between-country level of analysis, such that countries with generally higher levels of per capita reserves attain greater oil production than countries with generally lower levels of per capita reserves. Importantly, there is no support for the longitudinal form of the hypothesis. Regarding some of the other effects, we see that both ln(proven reserves) and conflict exhibit much more potent between-country than within-country effects. This is somewhat surprising since we have much more longitudinal information in the data (*T*=35) compared to cross-sectional variation (*N*=11). However, if there is little variation across time, estimates will not be as potent. And the data contain significant differences in the averages of these variables across countries, which contributes to the larger between-country effects.

[Figure 4 about here]

Senate Voting on Supreme Court Nominations, 1937-2005

The final substantive application involves a multilevel data analysis of Senate voting on Supreme Court nominations, where the dependent variable is binary (1=yea vote, 0=nay vote). Epstein et al. (2006) present an update of the Cameron, Cover, and Segal (CCS) (1990) model of Senate voting on nominees, which posits the influence of the following variables on a Senator's vote: a nominee's *lack of qualifications* (measured using content-analysis of newspaper editorials during the nomination process; ranges from 0 to 1, where higher values represent a less qualified nominee), whether the *president is in a strong political position* (president's party controls the Senate and president is not in fourth year of office), whether a senator is of the *same party as the president* (1=same party, 0=otherwise), and the *ideological distance* between the nominee and a senator. To measure ideological distance, Epstein et al. (2006, 299) employ a "bridging" procedure that uses the president's Poole-Rosenthal Common Space score in conjunction with the nominee's Segal-Cover (1989) ideological score to place senators and nominees in the same ideological space.[11] Epstein et al. take on some additional issues that I do not address here. On the whole, the authors find continued empirical support for the CCS model.

Epstein et al. include Senate votes on 40 nominations. There are 3,709 total votes. Treating this as a two-level hierarchical structure, the data consist of 3,709 votes nested within 40 nominations. Ideological distance and same party as the president are level-1 variables (varying across both Senate votes and nominations). Lack of qualifications and strong president are level-2 variables (varying only between nominations). Epstein et al. use a complete pooling approach (probit), which means that one cannot conclude with confidence whether the level-1 variables (ideological distance and same party) are within- or between-nomination effects. The model in Table 3 examines this issue. The left side of Table 3 presents a replication of Epstein et al.'s pooled probit model. The right side of the table includes a random intercept probit, which estimates within- and between-nomination effects of the level-1 variables as well as effects of the level-2 variables. It also presents tests of cluster confounding for the level-1 variables. In terms of model fit, a likelihood ratio test supports the specification of the random intercept model over

[11] The bridge is nominees chosen by a president who controls the Senate. See Epstein et al. (2006, 299) for more details on the measurement strategy.

a completely pooled approach. There is significant unobserved heterogeneity at the nomination level. The estimate of ρ indicates that 67% of the total error variance is accounted for by the nomination-level error.

[Table 3 about here]

Results from the random intercept model show that ideological distance exhibits a negative and statistically significant within-nomination effect on the probability of a yea vote and a positive but statistically insignificant between-nomination effect. Thus, Epstein et al.'s negative and significant pooled effect is driven by the strong within-nomination effect. While there is no statistically significant cluster confounding for ideological distance, the fact that the within and between effects are so drastically different strongly supports the need to distinguish the two types of effects. Thus, we can conclude that for a given nomination, as ideological distance between a senator and a nominee increases, the probability of a yea vote significantly decreases. Importantly, there is no contextual effect of ideological distance. That is, nominations for which there is a high average ideological distance do not have significantly different propensities of yea voting compared to nominations for which there is a low average ideological distance. For the effects of party, Epstein et al. find that senators of the same party as the president are significantly more likely to vote in favor of the nominee compared to senators not of the same party as the president. Results from the random intercept model show that the within-nomination effect of party is positive and statistically significant, while the between nomination effect is positive and marginally significant. Moreover, there is only marginal evidence of cluster confounding. For a given nomination, senators of the president's party are significantly more likely to vote yea than senators not of president's party. For the between effect of party, nominations in which the president has a high proportion of co-partisans in the Senate exhibit higher average probabilities of a yea vote compared to nominations where the President has a low proportion of co-partisans. For the remaining variables, we see that lack of qualifications has a negative and statistically significant effect, meaning that the more a nominee lacks qualifications, the less likely that nominee will receive a yea vote. While Epstein et al. found that the effect of strong president is positive and statistically significant, results from the random intercept model show that the effect of this variable is statistically insignificant.

Figure 5 presents graphical interpretations of the within- and between-cluster effects of ideological distance and same party as president. Akin to procedures used for producing Figures 3 and 4, the graphs plot the predicted probability of a yea vote while allowing the variable of interest to vary and holding remaining variables constant at their mean values.[12] For the within-nomination effects, I plot predictions when the between variable is set at the 10th percentile, median, and 90th percentile. Since this is a probit, note how the within-cluster effects are not parallel, since effect sizes will depend on the value of the between-nomination effect. Plot A shows the potent within-nomination effect of ideological distance. For values of distance that are nearly one unit away from the cluster mean, the probability of a yea vote approaches 0.4 for nominations where average distance is low. The null, and even slightly positive, between-nomination effect of ideological distance is displayed in plot B. This indicates that nominations

[12] The predicted probabilities are marginal with respect to the level-2 random effect. That is, u_{0j} is averaged over, as opposed to held constant at a particular value (i.e., conditional with respect to u_{0j}). Thus, these are akin to average partial effect (see Wooldridge 2002; Skrondal and Rabe-Hesketh 2004). In nonlinear models, whether predicted probabilities are marginal or conditional with respect to u_{0j} is an important distinction. In linear models, calculating predicted values of Y using each approach produces the same result.

with a high average distance evince roughly the same average probability of a yea vote compared to nominations with a low average distance. Plot C shows that for a given nomination with a low proportion of senators who are the president's co-partisans (the long-dashed line), the president's co-partisans are more likely to vote for the president's nominee than those not of the same party as the president. Note how the strength of this within-cluster effect dissipates slightly as the proportion of the Senate that is of the same party as the president increases. But then again, the overall propensity of a yea vote increases as well, which is also seen in plot D. For nominations in which only about 30% of the senate is of the same party as the president, the average probability of a yea vote is about 0.8, which is still quite high. But as this proportion increases to over 70%, the average probability of a yea vote increases to nearly 1.0.

[Figure 5 about here]

DISCUSSION AND CONCLUSION

As the substantive applications discussed above make clear, the modeling framework discussed in this paper has the potential to enrich both statistical analysis and substantive interpretations of effects in examinations of panel, TSCS, and multilevel data. Practitioners can be more explicit in communicating the substantive effects of certain variables by separating out the within- and between-cluster components of those effects. And the framework allows for statistical tests of whether these effects are statistically different. Given the substantive and statistical advantages of such a modeling strategy, analysts should be encouraged to implement many of these procedures.

Of course, this paper does not provide a panacea. There continue to be issues that need addressing in the analysis of clustered data. A primary issue in panel and TSCS data is dynamics. Since dynamics were not a focus of this paper, I adhered to a common and sensible practice to account for dynamics in panel and TSCS data: the inclusion of a lagged dependent variable (e.g., Beck and Katz 1996). As other work discusses, this may not always be the most optimal strategy, and analysts and methodologists should pay closer attention to issues of dynamics (e.g., Hsiao 2003; Wilson and Butler 2007). Moreover, analysts should not necessarily treat dynamics as a nuisance simply to be corrected. In panel and TSCS, dynamics are often of great substantive interest (e.g., Bartels, Box-Steffensmeier, Smidt, and Smith 2008; Green and Yoon 2002; Heckman 1981; Green, Palmquist, and Schickler 2002; Wawro 2002).

As I discussed in the "extensions" section in this paper, analysts should also pay attention to how a random coefficient model can produce substantively innovative tests of hypotheses. Inclusion of cross-level interactions can assess how contextual variables moderate lower-level effects. Also, akin to Gelman et al.'s (2006) multilevel example of income and vote choice, the RCM fosters innovative tests of how aggregate variation in *X* shapes the individual effect of that *X*. In the study of human rights abuse, for example, does variation in average GNP across countries moderate the longitudinal impact of GNP on rights abuse? In other words, will the within-country slopes in Figure 3C vary as a function of a country's average economic standing? Also, in the study Supreme Court nominations, one can envision how an RCM would help shed light on Epstein et al.'s contention that ideological distance has significantly increased over time, particularly since the nomination of Robert Bork. Specifying a random coefficient for the within-nomination operationalization of distance would allow the effect of this variable to vary across nominations and to retrieve comparable estimates of the effect of ideological distance across nominations. As seen by these and other examples, the use of this general multilevel modeling framework opens up new avenues for enhancing empirical analysis of panel, TSCS, and multilevel data.

REFERENCES

Bafumi, Joseph, and Andrew Gelman. 2006. “Fitting Multilevel Models When Predictors and Group Effects Correlate.” Presented at the Annual Meeting of the Midwest Political Science Association.

Bartels, Brandon L., Janet M. Box-Steffensmeier, Corwin D. Smidt, and Renee M. Smith. 2008. “Heterogeneity, State Dependence, and the Dynamics of Individual-Level Party Identification.” Typescript.

Beck, Nathaniel L. 2001. “Time-Series Cross-Section Data: What Have We Learned in the Past Few Years?” *Annual Review of Political Science* 4:271-93.

Beck, Nathaniel L., and Jonathan N. Katz. 1995. “What to Do (and Not to Do) with Time-Series—Cross-Section Data.” *American Political Science Review* 89: 634-647.

Beck, Nathaniel L., and Jonathan N. Katz. 1996. “Nuisance vs. Substance: Specifying and Estimating Time-Series—Cross-Section Models.” *Political Analysis* 6:1-36.

Beck, Nathaniel L., and Jonathan N. Katz. 2001. “Throwing Out the Baby with the Bathwater: A Comment on Green, Yoon and Kim.” *International Organization* 55:487-95.

Beck, Nathaniel L., and Jonathan N. Katz. 2007. “Random Coefficient Models for Time-Series—Cross-Section Data: Monte Carlo Experiments.” *Political Analysis* 15:182-195.

Baltagi, Badi H. 2005. *Econometric Analysis of Panel Data* (Third Edition). John Wiley & Sons.

Blaydes, Lisa. 2004. “Rewarding Impatience: A Bargaining and Enforcement Model of OPEC.” *International Organization* 58:213-37.

Blaydes, Lisa. 2006. “‘Rewarding Impatience’ Revisited: A Response to Goodrich.” *International Organization* 60:515-525.

Bowler, Shaun, Todd Donovan, and Robert Hanneman. 2003. “Art for Democracy’s Sake? Group Membership and Political Engagement in Europe.” *Journal of Politics* 65:1111-29.

Cameron, Charles M., Albert D. Cover, and Jeffrey A. Segal. 1990. “Senate Voting on Supreme Court Nominees: A Neoinstitutional Model.” *American Political Science Review* 84:525-34.

Epstein, Lee, Rene Lindstadt, Jeffrey A. Segal, and Chad Westerland. 2006. “The Changing Dynamics of Senate Voting on Supreme Court Nominees.” *Journal of Politics* 68:296-307.

Gelman, Andrew, Boris Shor, Joseph Bafumi, and David Park. 2006. “Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut?" *Quarterly Journal of Political Science* 2:345-67.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Goodrich, Ben. 2006. "A Comment on 'Rewarding Impatience.'" *International Organization* 60:499-513.

Green, Donald P., Soo Yeon Kim, and David Yoon. 2001. "Dirty Pool." *International Organization*. 55:441-68.

Green, Donald, Bradley Palmquist, and Eric Schickler. 2002. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. New Haven: Yale University Press.

Green, Donald P., David H. Yoon. 2002. "Reconciling Individual and Aggregate Evidence Concerning Partisan Stability: Applying Time-Series Models to Panel Survey Data." *Political Analysis* 10:1-24.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46:1251-71.

Heckman, James J. 1981a. "Heterogeneity and State Dependence." In *Studies in Labor Markets*, ed. S. Rosen. Chicago, IL: University of Chicago Press.

Hsiao, Cheng. 2003. *Analysis of Panel Data*. New York: Cambridge University Press.

Kristensen, Ida P., and Gregory Wawro. 2003. "Lagging the Dog? The Robustness of Panel Corrected Standard Errors in the Presence of Serial Correlation and Observation Specific Effects." Presented at the Political Methodology Conference.

Martin, Andrew D. 2001. "Congressional Decision Making and the Separation of Powers." *American Political Science Review* 95:361-378.

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

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* 88:853-72.

Poe, Steven C., C. Neal Tate, and Linda Camp Keith. 1999. "Repression of the Human Right to Personal Integrity Revisited: A Global Cross-National Study Covering the Years 1976-1993." *International Studies Quarterly* 43:291-313.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2005. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models*. Thousand Oaks: Sage.

Rodriguez, German, and Noreen Goldman. 1995. "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses." *Journal of the Royal Statistical Society, Series A* 158:73-89.

Rodriguez, German, and Noreen Goldman. 2001. "Improved Estimation Procedures for Multilevel Models with Binary Response: A Case-Study." *Journal of the Royal Statistical Society, Series A* 164:339-55.

Shor, Boris, Joseph Bafumi, Luke Keele, and David Park. 2007. "A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data." *Political Analysis* 15:165-81.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall.

Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46:218-37.

Stimson, James A. 1985. "Regression in Space and Time: A Statistical Essay." *American Journal of Political Science* 29:914-47.

Wawro, Gregory. 2002. "Estimating Dynamic Panel Models in Political Science." *Political Analysis* 10:25–48.

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

Wilson, Sven E., and Daniel M. Butler. 2007. "A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications." *Political Analysis* 15:101-23.

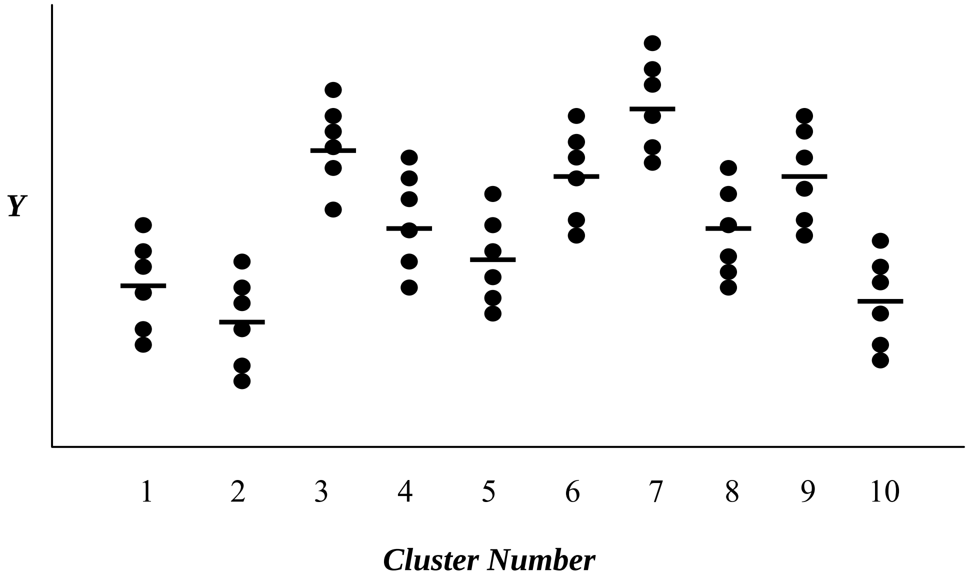
Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Zeger, Scott L., and Kung-Yee Liang. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." *Biometrics* 42:121-30.

Zorn, Christopher J.W. 2001a. "Generalized Estimating Equation Models for Correlated Data: A Review With Applications." *American Journal of Political Science* 45:470-90.

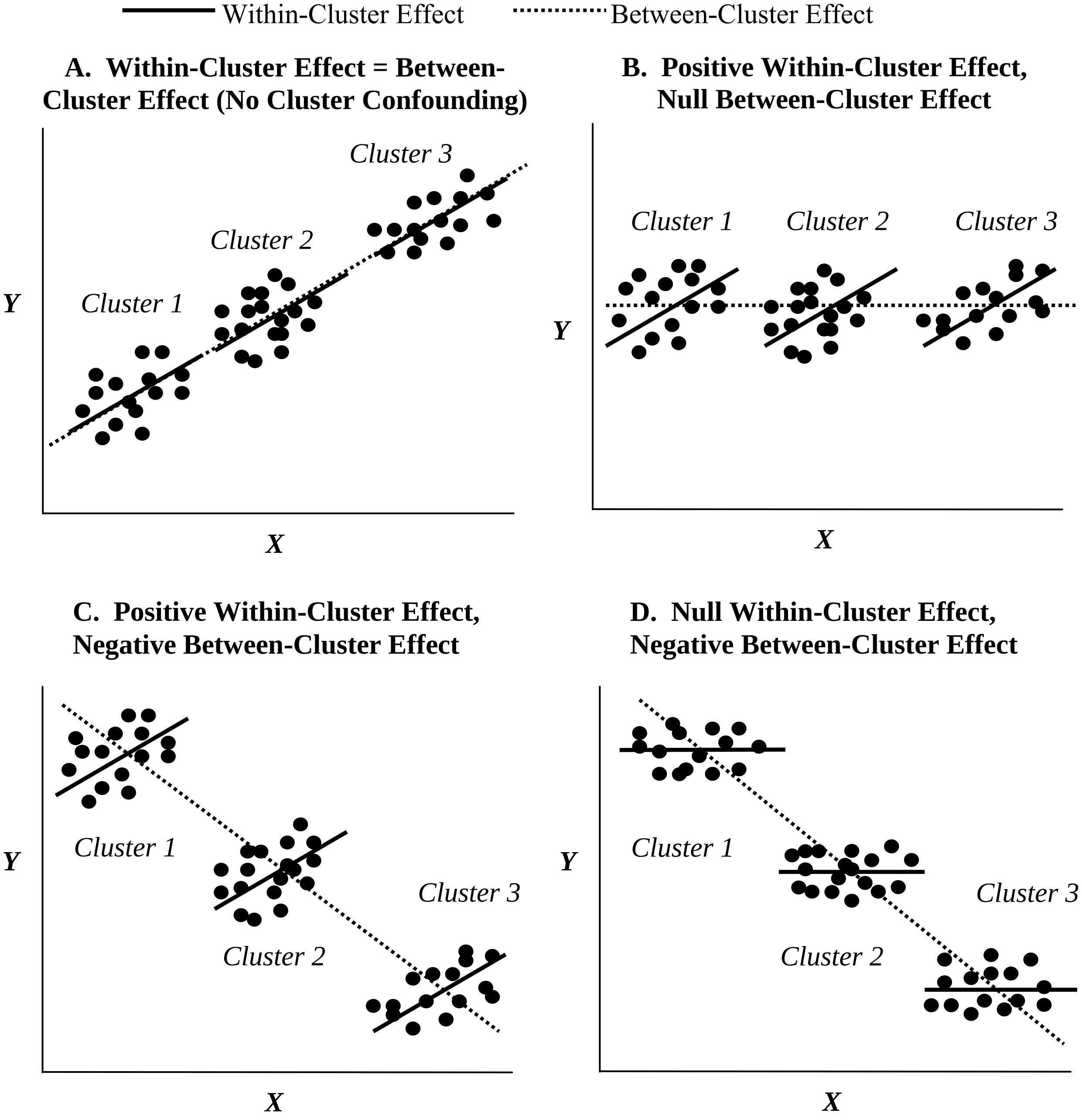
Zorn, Christopher. 2001b. "Estimating Between- and Within-Cluster Covariate Effect, with an Application to Models of International Disputes." *International Interactions* 27:433-45.

Figure 1: Illustration of Unobserved Heterogeneity Across Clusters



Note: Dots represent responses within a given cluster. Dashes represent the means of *Y* for each cluster.

Figure 2: Illustration of Cluster Confounding



A. Within-Cluster Effect = Between-Cluster Effect (No Cluster Confounding)

B. Positive Within-Cluster Effect, Null Between-Cluster Effect

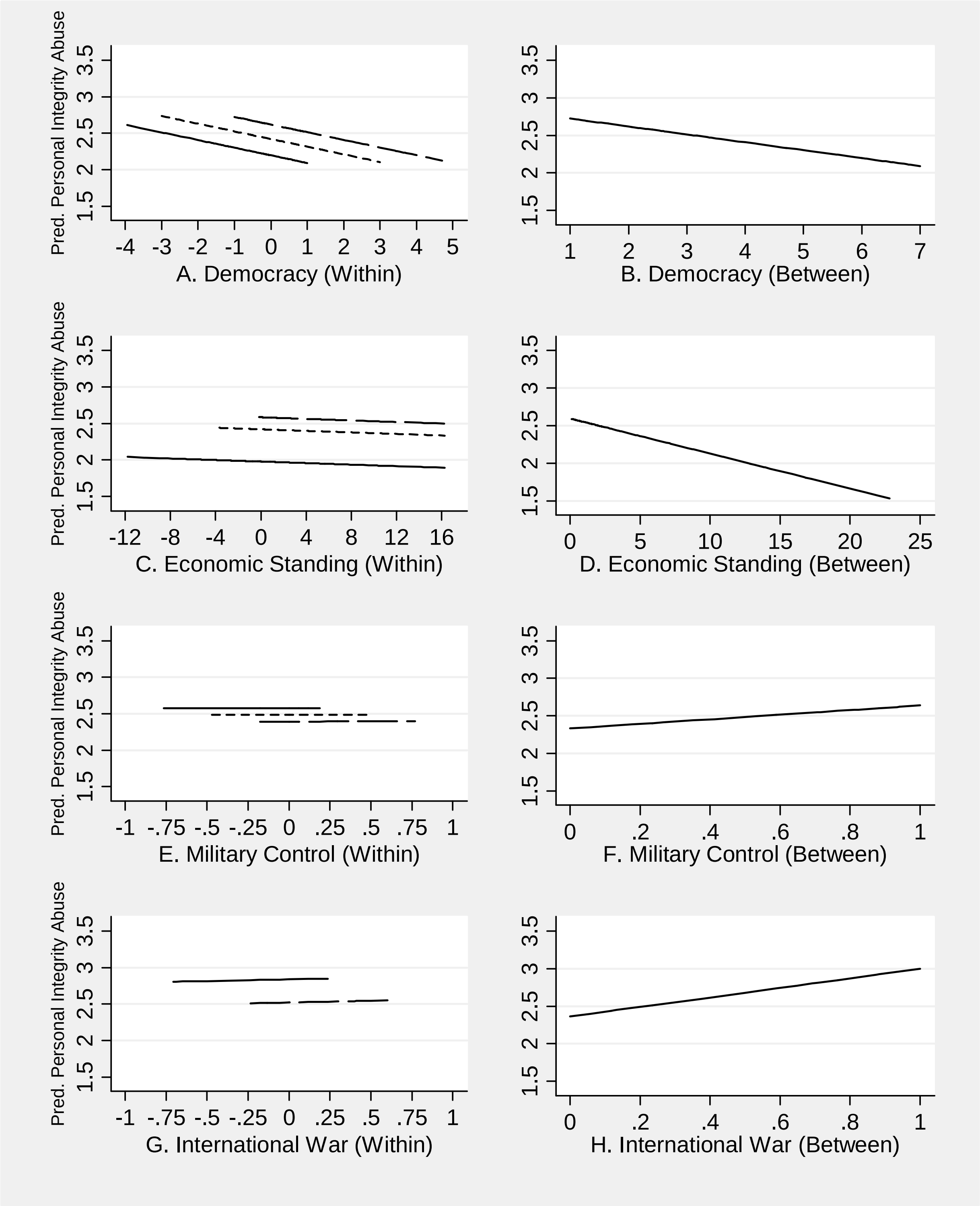
C. Positive Within-Cluster Effect, Negative Between-Cluster Effect

D. Null Within-Cluster Effect, Negative Between-Cluster Effect

Table 1: Models of Global Human Rights Abuse – Amnesty International Models, 1977-1993

				Linear Random Intercept Model (Maximum Likelihood)								
	Poe et al. (1999) OLS Results			**Within-Country Effects**			**Between-Country Effects**			**Abs(Within - Between)**		
	Coef.	**(PCSE)**	**p**	**Coef.**	**(SE)**	**p**	**Coef.**	**(SE)**	**p**	**Coef.**	**(SE)**	**p**
Rights Abuse$_{t\text{-}1}$	0.65	(0.02)	0.00	0.38	(0.02)	0.00	-	-	-	-	-	-
Democracy (Freedom House)	-0.06	(0.01)	0.00	-0.10	(0.01)	0.00	-0.11	(0.03)	0.00	0.00	(0.04)	0.99
Population Size	0.07	(0.01)	0.00	0.22	(0.09)	0.02	0.18	(0.02)	0.00	0.04	(0.10)	0.66
Population Change	0.00	(0.00)	0.23	0.00	(0.00)	0.70	0.07	(0.03)	0.03	0.07	(0.03)	0.03
Economic Standing	-0.02	(0.00)	0.00	-0.01	(0.00)	0.28	-0.05	(0.01)	0.00	0.04	(0.01)	0.00
% Economic Change	0.00	(0.00)	0.35	0.00	(0.00)	0.22	0.00	(0.01)	0.79	0.00	(0.01)	0.89
Leftist Government	-0.17	(0.04)	0.00	-0.04	(0.08)	0.59	-0.32	(0.15)	0.04	0.28	(0.17)	0.10
Military Control	0.09	(0.03)	0.01	0.00	(0.05)	0.97	0.31	(0.12)	0.01	0.31	(0.13)	0.02
British Cultural Influence	-0.08	(0.03)	0.00	-	-	-	-0.23	(0.09)	0.01	-	-	-
International War	0.14	(0.04)	0.00	0.05	(0.06)	0.42	0.63	(0.21)	0.00	0.59	(0.22)	0.01
Civil War	0.50	(0.05)	0.00	0.46	(0.06)	0.00	1.53	(0.19)	0.00	1.06	(0.20)	0.00
Constant	0.07	(0.10)	0.50	-0.13	(0.42)	0.75						
Observations	N=164, T(avg.)=15.1 Tot. Obs.=2,471			N=164, T(avg.)=15.1 Tot. Obs.=2,471								
Model χ^2	8911.44, p<.001			832.83, p<.001								
Var(Level-1 Error)	-			0.29								
Var(Level-2 Error)	-			0.25								
ρ (Level-2 Error / Total Error)	-			0.46								
LR Test (H_0: Level-2 Error=0)	-			χ^2=1029.94, p=<.001								

Figure 3: Within-Country and Between-Country Effects of Selected Variables on Personal Integrity Abuse



Note: For within-country effects (A, C, E, and G): ——— High value of between-country variable
- - - - - - Medium value — — — — Low value; see text for further details.

Table 2: Models of Oil Production for OPEC Countries, 1960-1995

					Linear Random Intercept Model (Maximum Likelihood)					
	Blaydes Model 3 with PCSEs		**Blaydes Model 3 with Random Effects (FGLS)**		**Within-Country Effects**		**Between-Country Effects**		**Abs(Within-Between)**	
	Coef. (PCSE)	**p**	**Coef. (SE)**	**p**	**Coef. (SE)**	**p**	**Coef. (SE)**	**p**	**Coef. (SE)**	**p**
Ln(Proven Reserves)	0.19	0.00	0.18	0.00	0.09	0.03	0.76	0.00	0.67	0.00
	(0.02)		(0.02)		(0.04)		(0.04)		(0.06)	
Ln(Per Capita Reserves)	-0.02	0.01	-0.02	0.05	0.05	0.20	-0.07	0.00	0.12	0.01
	(0.01)		(0.01)		(0.04)		(0.02)		(0.04)	
Ln(Per Capita Reserves Squared)	-0.01	0.01	0.00	0.13	0.00	0.93	-0.04	0.00	0.04	0.00
	(0.00)		(0.00)		(0.01)		(0.01)		(0.01)	
One-Year Lag in Crude Oil Production	0.70	0.00	0.70	0.00	0.71	0.00	-	-	-	-
	(0.02)		(0.02)		(0.02)					
Conflict	-0.42	0.00	-0.40	0.00	-0.38	0.00	-1.57	0.00	1.19	0.00
	(0.07)		(0.06)		(0.07)		(0.33)		(0.34)	
Constant	0.37	0.00	0.46	0.01	0.05	0.91				
	(0.11)		(0.17)		(0.42)					
Observations	N=11, T=35 Tot. Obs.=385		N=11, T=35 Tot. Obs.=385		N=11, T=35 Tot. Obs.=385					
Model χ^2	5736.82, p<.001		4193.97, p<.001		836.56, p<.001					
Var(Level-1 Error)	-		0.07		0.07					
Var(Level-2 Error)	-		0.001		0.01					
ρ (Level-2 Error Var / Total ErrorVar)	-		0.02		0.17					
LR Test (H$_0$: Level-2 Error=0)	-		χ^2=1.30, p=0.25		χ^2=46.42, p<.001					

Figure 4: Within-Country and Between-Country Effect of Per Capita Reserves on Crude Oil Production

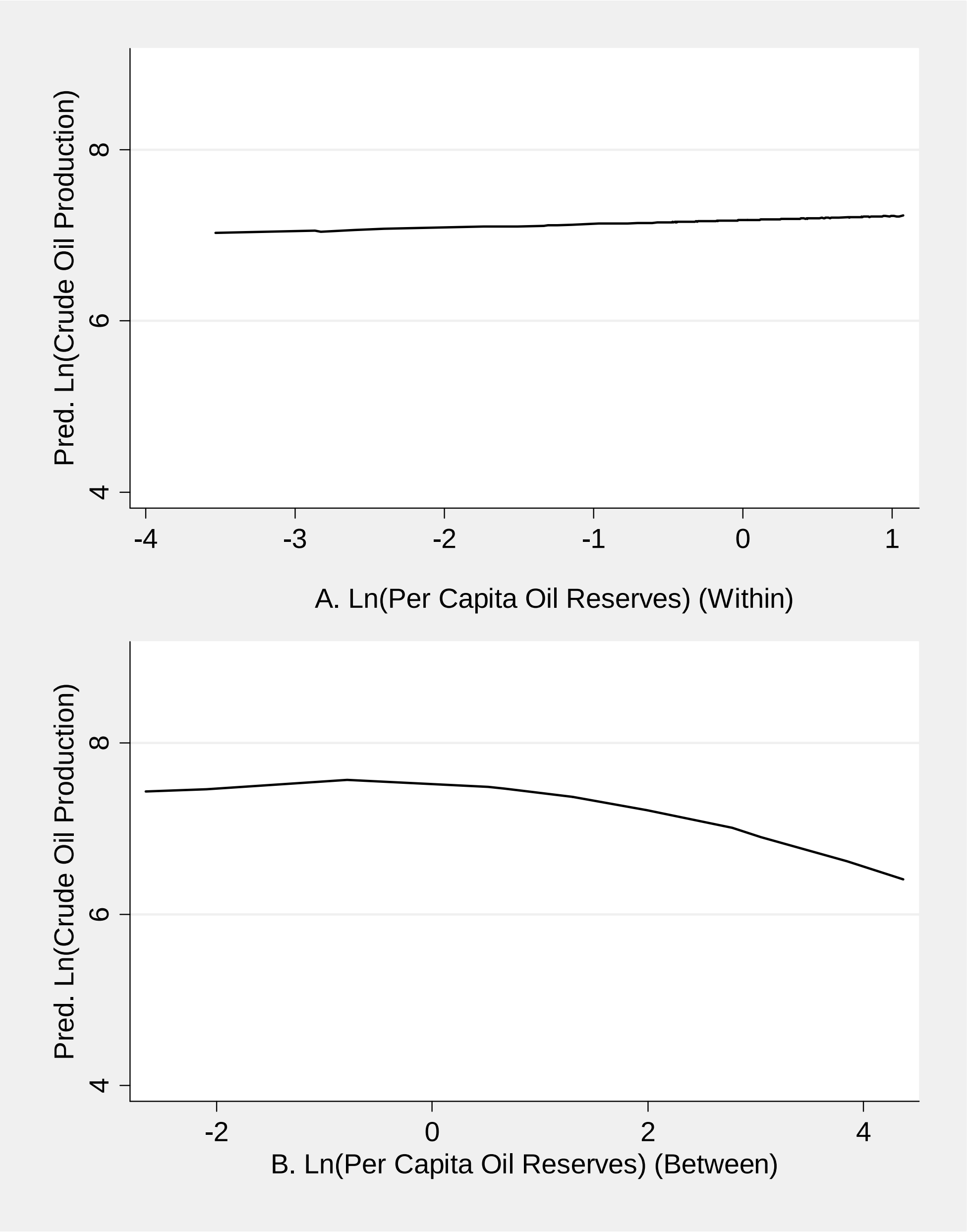
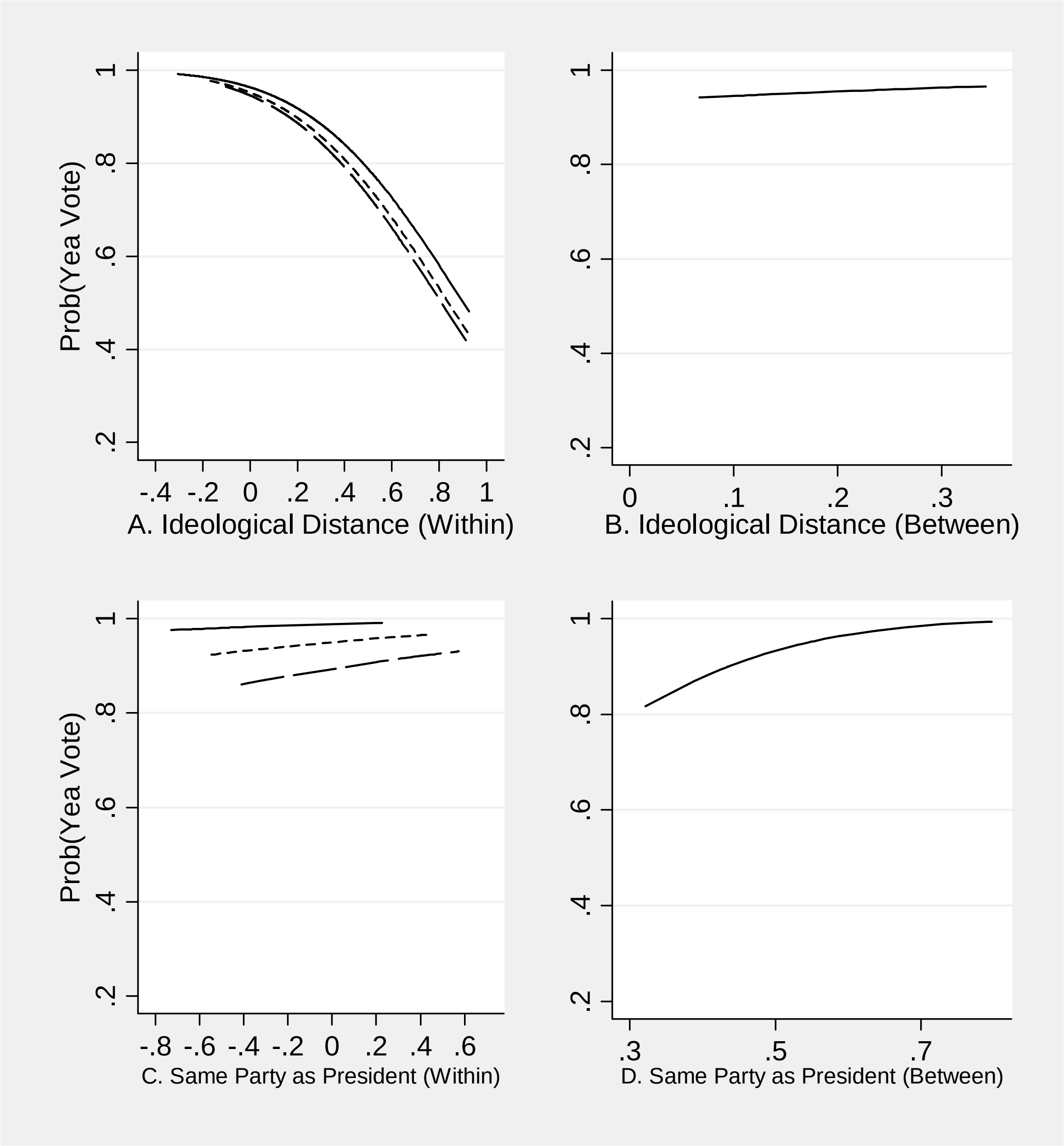


Table 3: Models of Senate Voting on Supreme Court Nominations

			Random Intercept Probit Model					
	Epstein et al. Probit Model		**Within-Nomination Effects**		**Between-Nomination Effects**		**Abs(Within-Between)**	
	Coef. (SE)	**p**	**Coef. (SE)**	**p**	**Coef. (SE)**	**p**	**Coef. (SE)**	**p**
Ideological Distance	-2.24 (0.14)	0.00	-3.48 (0.24)	0.00	1.58 (3.74)	0.67	5.06 (3.76)	0.18
Same Party	0.71 (0.08)	0.00	0.70 (0.10)	0.00	5.81 (3.15)	0.07	5.11 (3.16)	0.11
Lack of Qualifications	-2.32 (0.12)	0.00	-	-	-4.69 (0.97)	0.00	-	-
Strong President	0.77 (0.07)	0.00	-	-	0.60 (0.69)	0.38	-	-
Constant	1.82 (0.08)	0.00	0.12 (1.79)	0.95				
Observations	N=3,709		Level-1 units (votes): 3,709 Level-2 units (nominations): 40					
Model χ^2	581.42, p<.001		381.37, p<.001					
Var(Level-2 Error)	-		2.04					
ρ (Level-2 Error / Total Error)	-		0.67					
LR Test (H_0: Level-2 Error=0)	-		χ^2=446.59, p<.001					

Figure 5: Within-Nomination and Between-Nomination Effects of Ideological Distance and Party on Senate Voting for Supreme Court Nominations



Note: For within-nomination effects (A and C): ——— High value of between-country variable (90th pctile) ------ Median value — — — —Low value (10th pctile); see text for further details.