MODELING MULTILEVEL DATA STRUCTURES*

Bradford S. Jones
University of Arizona†

Marco R. Steenbergen
University of North Carolina, Chapel Hill

***Abstract**: Although integrating multiple levels of data into an analysis can often yield better inferences about the phenomenon under study, traditional methodologies used to combine multiple levels of data are problematic. In this paper, we discuss several methodologies under the rubric of multilevel analysis. Multilevel methods, we argue, provide researchers, particularly researchers using comparative data, substantial leverage in overcoming the typical problems associated with either ignoring multiple levels of data, or problems associated with combining lower-level and higher-level data (including overcoming implicit assumptions of fixed and constant effects). The paper discusses several variants of the multilevel model and provides an application of individual-level support for European integration using comparative political data from Western Europe.*

***Keywords:** multilevel analysis, random coefficients models, contextual analysis, comparative politics*

INTRODUCTION

Many problems in political science can be studied at multiple levels of analysis and combining these levels into a single analytical approach is often very desirable. A considerable number of theories and hypotheses in political science hinge on the presumption that "something" observed at one level affects or is related to "something" observed at another level. Yet despite the prevalence of cross-level or multilevel theories and hypotheses of political behavior, political scientists have been slow to adopt statistical methods developed for analyzing multilevel data structures.[1] The goal of this paper is to describe and illustrate these methods for problems of comparative analysis. We take a very broad view of "comparative analysis" in this paper. Any research design that generates inferences explicitly based on comparisons across political "units" suffices to be comparative analysis. A political unit may be geographically defined (i.e. states or countries), temporally defined (i.e. comparisons of elections across time), or socially defined (for example, political or social groups, class, etc.).

We choose comparative analysis as our domain of application because multilevel data structures are prevalent in this type of research. Indeed, some have even made it a defining characteristic of comparative research that multiple levels of analysis are

* Paper prepared for the 14th annual meeting of the Political Methodology Society, Columbus, OH, July 25, 1997. The authors would like to thank Thad Brown, Suzy DeBoef, Brian Crisp, Bill Dixon, Ita Kreft, Bill Mishler, George Rabinowitz, Leonard Ray, and Michael Sobel for helpful comments, suggestions, and insights.

† The authors have been listed in alphabetic order.

analyzed simultaneously (e.g. Rokkan 1966; Przeworski and Teune 1970; but see Ragin 1987, 4). The methods we consider are known under a variety of names – multilevel analysis, hierarchical models, random coefficients models, and variance components analysis. The common element of all of these methods is that a dependent variable is analyzed at the lowest level of analysis in which a researcher is interested. This variable is analyzed as a function of predictors measured at this level of analysis and of predictors measured at one or more higher levels of analysis. Moreover, the impact of the predictors at the lowest level of analysis is allowed to randomly vary over the higher levels of analysis.

Our strategy in this paper is to first outline the motivation for conducting multilevel analysis. Second, we discuss some statistical problems inherent with multilevel data structures and consider why traditional approaches for dealing with these kinds of structures are problematic. Third, we outline the multilevel model and describe how it helps alleviate some of the problems associated with multilevel data structures. Fourth, we discuss the statistical aspects of multilevel analysis, including a consideration of interpretation, modeling strategies, and software issues. Fifth, we present applications of multilevel techniques. And sixth, we conclude with a discussion of some caveats and pitfalls associated with multilevel methods.

MOTIVATION FOR MULTILEVEL MODELS

The motivation for multilevel modeling lies in the assumption that variation in a dependent variable is a function of both lower-level and higher-level factors. Furthermore, the relationship between these factors and the dependent variable is not assumed to be fixed or constant across space or time. Therefore, when examining individual-level data, variation in behavior (or attitudes, preferences, and so forth) is not only a function of individual-level attributes, but also extra-individual factors or more generally, macro-level factors.[2] From an econometric point-of-view, this implies regression coefficients in micro-level models are not fixed, but allowed to vary across these factors. What "these factors" are, or course, is a theoretical question. In this section, we consider various theoretical and practical motivations for combining multiple levels of data.

Cross-Area Comparative Analysis. We use the term "cross-area" to denote research designs that comparatively examine multiple geographical "units." The units in this kind of design may involve countries, geographical regions that extend beyond national borders, or regions within a single country. Despite the unit of analysis, a perennial concern in cross-area comparative political science is the issue of "contextual variation" (c.f. Ragin 1987; Collier 1993; Agnew 1987, 1996a). Unfortunately, what actually comprises "context" is often ill-defined or generally cast in terms of amorphous "political culture" arguments. Furthermore, the issue of actually being able to model "context" has been hotly debated among comparative methods scholars for many years (for example, Kalleberg 1966; Rokkan 1966, 1971; Sartori 1970, 1991; Przeworski and Teune 1970; Lijphardt 1971; Geertz 1973; Agnew 1987, 1996a; Ragin 1987; King, Keohane, and Verba 1994; King 1996, to name a few). The cleavages of this debate are too complex to fully document here; however, one aspect of the "contextual problem" has been the

countervailing view of the necessity and ability of researchers to engage in quantitative analysis of cross-area data.

Political contexts, some have argued, are too complex, too varied, and too nuanced to be adequately captured in econometric models. Instead, "thick description" (Geertz 1973) or single case-study approaches are the only valid means toward comparative analysis, at least from this perspective. Nevertheless, that political contexts vary, have led some comparativists to actually *advocate* "large-n" quantitative analyses (for example, Przeworski and Teune 1970 and more generally, Jackman 1985). The argument here, roughly put, is that in order to understand the importance of contextual variation, one actually needs variation in contexts. Expressed in this way, the problem of modeling contextual variation is akin to the "case selection" problem delineated generally by Achen (1986) and in terms of comparative analysis, by Geddes (1990) and King, Keohane, and Verba (1994).[3]

But to "select cases" implies there are cases to select. The argument is frequently made that comparable cross-area data are rarely available for comparative analysis, and furthermore, the data available are largely aggregated, country-level data (see Collier 1993, for an overview of these concerns). So the question becomes, how does one model "context"[4] when one, apparently, has few data points (and the few available are aggregated)? The answer to this question leads to circularity. Because of the inherent problems with cross-area data (lack of it, incomparability), thick description or single-case studies are, by many arguments, the only valid modes of analysis. But then (as noted above) contextual variation cannot be modeled because there is no context that varies in single case studies. Therefore, "large-n" analyses need to be performed to capture this variation. *But* comparable data are rarely available… . And so on.

This "data problem" elicits both practical and theoretical problems. Practically, the lack of extensive aggregate *and* individual-level data precludes, in some instances, many quantitative methodologies.[5] This is particularly problematic for researchers who attempt to model cross-area variation in individual-level behavior *and* simultaneously try to account or "control" for contextual effects. Indeed, individual-level analyses have been problematic because of the preponderance of aggregated data, to the exclusion of individual-level data. Nevertheless, two developments, one methodological, the other data-related, have made inference-making at the individual-level possible.

First, King's (1997) work on the ecological inference problem seems to provide an avenue for comparative researchers to generate individual-level inferences from aggregated data. Because there is a relative wealth of aggregate data (when compared to individual-level data) across countries and regions, King's solution to the ecological inference problem may elicit more attempts to understand variation in individual-level behavior.[6] The approach we take in this paper differs from King's work (although aspects of statistical estimation are similar) because we presume the existence of *both* individual-level *and* aggregate-level data.

Fortuitously, the second development in cross-area analysis has been the emergence of individual-level survey data.[7] Although the *Euro-barometer* has been around for many years, the *World Values Survey* (Basanez, Inglehart, and Moreno 1997) promises to add a considerably wider range of individual-level data for many global regions. Additionally, a host of other regional surveys (see analyses of these data in Gibson 1996, Gibson and Duch 1992, Gibson, Duch, and Tedin 1992, Mishler and Rose 1997, and so on) indicates

the problem with limited individual-level data is dissipating rapidly.[8] Recent research using cross-area individual-level data suggests substantial leverage may be gained in understanding processes of citizen and elite opinion dynamics, support for democracy, racial and ethnic tolerance, and so forth (see Franklin, Marsh, and McLaren 1994; Franklin and Rudig 1995; Franklin, Van Der Eijk , and Marsh 1995; Gibson and Duch 1992; Gibson, Duch, and Tedin 1992; Gibson and Caldiera 1996; Gibson 1996; Mishler and Rose 1997 for very recent examples of this work).

With the emergence of individual-level data, the theoretical "data problem" then becomes one of relating individual-level data to aggregate-level data. We think the concerns comparativists have with contextual variation, and more generally, with the relationship between macro-level factors and individual-level factors can be addressed with the multilevel techniques we discuss in this paper. The argument that contextual variation precludes systematic quantitative analyses of individual-level behavior in cross-area research, we believe, is now largely vacuous. The growing body of individual-level and aggregate-level data in comparative politics permits estimation of models that can combine data measured at different levels. Problems of heterogeneity, assumptions of fixed effects, and most generally, contextual variation, can be accounted for with multilevel techniques.

Pooled Time-Series Cross-Sections. Comparative political data are frequently analyzed as pooled time-series cross-sections. Work in the political methodology literature has extensively considered the special problems that emanate from such data (c.f. Beck 1983; Stimson 1985; Beck and Katz 1995, 1996a, 1996b). Recently, Beck and Katz (1996b) and Western (1997) have considered estimation of random coefficients models for pooled time-series cross-section designs. Among the statistical problems that emerge from such designs is what Western (1997) calls "causal heterogeneity."

For example, if one is interested in the relationship between some set of covariates and economic conditions (such as unemployment; see Western 1997), unaccounted-for causal heterogeneity may lead to incorrect or imprecise inferences. As he notes, only if one assumes the relationship between covariates and the dependent variable is constant across countries does one need *not* worry about causal heterogeneity. Unfortunately, given the pronounced relationship of "contextual factors" that vary across countries (Przeworski and Teune 1970), it is unlikely that the same forces operating in one country are constant across all countries (Western 1997). In time series analysis, this suggests that "fixed" features of a country (for example, institutional factors that are largely time-invariant) may induce heterogeneity because parameters in standard time series are agnostic to country-specific factors that induce variable coefficients. This kind of heterogeneity is therefore left unaccounted-for and relegated to the error structure. But if institutions are "nested" within countries then parameter estimates in the time-series may vary in accordance with institutional or contextual variation (Western 1997). The models discussed here and in Western (1997) provide researchers who work with pooled time-series cross-sections some leverage in accounting for this kind of heterogeneity. Thus, this type of design in comparative politics provides another theoretical motivation for using multilevel analysis.

In addition to pooled cross-section time-series designs for aggregate comparative data, multilevel methods may prove useful for pooled designs with individual-level data.

For example, "election effects" may be modeled by thinking of individuals as being nested within campaigns or elections. National or regional aggregate political factors may produce varying coefficients for individual-level models, *if* individuals respond or behave differently to changing national conditions (c.f. Kramer 1983; Haller and Norpoth 1994) or political campaigns (c.f. Kahn and Kenney 1997, Westlye 1994). More generally, comparative analyses of elections over time, or comparative analyses of multiple campaigns within a single national election may require examining data measured at multiple levels. In either design, multilevel methods may be appropriate tools.

Comparative Institutions and Legislative Behavior. Scholars interested in legislative behavior frequently deal with data measured at multiple levels. For example, data on voting records of U.S. House and Senate members is widely available and an abundance of research suggests these voting records vary, at least to some degree, on constituency characteristics and institutional attributes[9] (c.f. Kingdon 1992; Jackson and King 1989; Box-Steffensmeier, Arnold and Zorn 1997). Analyses of legislative behavior outside the United States has also found considerable variation in legislator behavior attributable to constituency characteristics (c.f. Ames 1987, 1995). And more generally, legislator behavior across a variety of activities (voting, committee selection, electoral behavior, etc.) has been shown to be related to institutional characteristics as well as preferences of constituencies (c.f. Hibbing 1992; Katz and Sala 1995).

Comparative legislative research has focused on, among other things, the extent to which legislators pursue the "personal vote" (Cain, Ferejohn, and Fiorina 1987; Carey and Shugart 1995). Carey and Shugart (1995) hypothesize that legislator behavior is substantially conditional upon nomination processes and electoral laws. Such facets of a country's electoral system clearly vary across countries and how this macro-level variation combines with legislator characteristics suggests that a model combining both the macro- and micro-level data is appropriate.

Contextual Analysis. Contextual analysis of political behavior is a research field where the assumption of aggregated influence on individual-level opinions and behavior is most explicitly made (see Huckfeldt and Sprague 1993). The major supposition of contextual analysis is that the "contextual effect…arises due to social interaction within an environment" (Huckfeldt and Sprague 1993, 289). This environment (i.e. the context) may be spatially defined, for example, in terms of local neighborhoods (c.f. Huckfeldt and Sprague 1987; Brown 1981, 1988) or in terms of local "social networks" (c.f. McKuen and Brown 1987). Thus, requisite data for contextual analysis involves information gathered both at the individual level and at the extra-individual level.

Apart from spatial or geographic definitions of context, political scientists have long regarded political and social groupings as a source of contextual variation (c.f. Uhlaner 1989; Lau 1990; Smith 1990). The basic idea here is that group membership, or more specifically, attributes of the group itself, exert influence on an individual's opinions, attitudes, preferences, or behaviors. Thus, when treated as a "contextual effect," individual-level outcomes are conditioned on not only individual factors, but also group-level effects.[10] And the social influence of groups need not be confined to tangible affiliations (for example, membership in the National Rifle Association or the Sierra

Club). The notion of social identification (c.f. Tajfel 1978) suggests that individuals may "identify" with many social groupings that are not necessarily well defined or bounded as is the case with affiliational groups. For example, individuals may identify with ethnic, racial, socioeconomic, or class-based groupings. In this case, the contextual "unit" is very disperse but the contextual "effect" still implies that a macro or group level "consciousness" or awareness influences individual-level judgements and behavior.[11]

So broadly defined, contextual analysis provides a convenient segue for multilevel modeling.[12] Contextual theories or hypotheses posit that individual behavior is some function of both individual-level and extra-individual factors, and therefore, data at multiple levels need to be considered jointly. How these multiple levels of data are combined has been an on-going issue in contextual analysis (c.f. Boyd and Iverson 1979; Iverson 1991; Sprague 1976, 1982; Sitpak and Henslar 1982) and we argue the theoretical underpinnings of contextual analysis naturally leads to a consideration of multilevel methods.

To conclude this section, we have delineated several theoretical, practical, and substantive motivations that provide an avenue toward multilevel modeling. Clearly, there are more technically oriented reasons why one might consider analyzing multiple levels of data, and of course, we address these issues in detail below; however, from an applied perspective, there are numerous hypotheses and theories in political science that leads us to consider combining lower and higher levels of data. In the next section, we discuss why traditional methods of combining multilevel data are problematic.

COMBINING MULTIPLE LEVELS OF DATA

Comparative research frequently involves combining multiple levels of analysis; however, many standard techniques for combining data are inadequate. In this section, we consider the problems associated with the "separate models" approach, the dummy variables approach, and most generally, the interactive model approach.

Separate Models. One method to "combine" multiple levels of data in a research design is to *avoid* combination, or to eliminate variance in higher-level factors. For example, a cross-area research design may consider how individuals nested within a Western European country support their country's membership in the European Union. Because "contextual," historical, or other extra-individual factors may influence in some way, individual-level attitudes and preferences, these factors are "held constant" by estimating separate individual-level models for every country:

$$
\begin{aligned}
y_{i1} &= \beta_0 + \beta_1 x_i + \varepsilon_i \\
y_{i2} &= \alpha_0 + \alpha_1 x_i + \varepsilon_i \\
&\vdots \\
y_{ij} &= \omega_0 + \omega_1 x_i + \varepsilon_i
\end{aligned}
$$

In this case, the parameters reflect the relationship between x_i and the response variable.

The different Greek characters illustrate the specific parameters derived from estimation of separate models for each of the *j* countries. Such a design can avoid combining cross-level data because country-specific factors (or contextual effects) are essentially held constant because only individual-level data nested *within* the country are used to derive parameter estimates. In this sense, context is implicitly modeled by *not* being modeled. Frequently in such designs, "eyeball" comparisons are made across the rows of coefficients estimated for each country and assessment (often times nonstatistical assessment) is made by comparing and contrasting differences in magnitudes of the coefficient estimates. With respect to understanding how macro-level factors relate to or interact with lower-level factors, however, this design is problematic.

Because contextual factors that may vary across the *j* countries are ignored (a result of the separate models), it is difficult to discern the relationship between macro and micro-level variables. Although coefficient estimates may (or may not) vary across the separate-country regressions, typical "eyeball" tests do not provide sound evidence of statistically significant variation (or lack thereof). And while statistical comparisons across pairs of regressions (for sets of coefficients) are possible using multiple Chow tests,[13] it is not obvious what the import of these tests are in terms of making inferences about specific parameters (Bartels 1996).[14] Eyeball comparisons or Chow test statistics only provide information about differences across sets of equations, and therefore provide no information on *why* the variation exists in the data across the *j* units.

Differences in parameter estimates may be attributable to unobserved heterogeneity.[15] The source of this heterogeneity may center on contextual factors or some other type of unobservable factor that exhibits influence on individual-level data. Ironically enough, it is because of this heterogeneity—the unobserved (or unmeasured contextual factors)—that some comparativists resort to disaggregating data by country (or some other unit) rather than pooling observations. Indeed if the disturbance variances across the countries (or more generally, units) are unequal, such that $E(\sigma_i^2) \neq E(\sigma_j^2)$, then an appropriate modeling strategy is to disaggregate the data and estimate models on subsamples (c.f. Greene 1993, 236;[16] but see Bartels 1996), or estimate models that can accommodate heteroskedastic error structures.

The problem with not pooling the data, at least with regard to making multilevel inferences, is that a considerable amount of information is lost, wasted, or ignored by failing to pool the observations. And even more problematic, the influence of contextual factors—influence commonly hypothesized in comparative research—cannot be assessed through disaggregation. Thus, if the theory suggests a combination of multiple levels of data, then unit-specific models fail to capture the theory because they are incapable of discerning the relationship between higher and lower levels of data.

Dummy Variables Models. One technique often used to gain some leverage on the heterogeneity problem discussed above is through the use of dummy variables. Because data pooled across contextual units potentially elicit heterogeneity, dummy variables are commonly used to "capture" this heterogeneity on the right-hand side of the equation. Capturing heterogeneity through dummy variables may involve inclusion of separate indicators for the *j*-1 contextual units:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_1 + \beta_3 D_2 + ... + \beta_k D_{j-1} + \varepsilon_i,$$

where the *D* terms represent binary indicators for the *j-1* units. The dummy variables approach is frequently used with pooled time-series data in the form of least squares dummy variables models (LSDV) to model either "space" or "time"[17] problems (c.f. Stimson 1985, Sayrs 1989, Hardy 1993) and in comparative analyses to model unit-specific heterogeneity. In this sense, inclusion of dummy variables acts as a control, of sorts, for the "noise" inherent in the pooled cross-sections, or in the time dynamic. And to that end, the dummy variables approach is perfectly reasonable and justifiable.[18] However, if the intent of the model is to derive substantive inferences about the relationship between multiple levels of data, then the dummy variables approach is problematic.

To the extent one is concerned with aggregate-level influence, a dummy variable indicator for a contextual unit provides sparse information about macro-micro relationships. With the use of dummy variables, the researcher is implicitly arguing that important differences exist between contextual units, but can say very little about the mechanisms eliciting these differences because dummy variables do not represent anything "substantive."[19] Little insight is gained from examining parameter estimates of dummy variables, except in noting that some unit-specific influence is at work And given the bluntness of the information yielded by dummy variables, the inference problem gets worse as the number of contextual units increases. Pooling data generated across a number of units induces a "proliferation of parameters" problem. If one is serious about modeling unit-specific factors, then one will necessarily need to include a substantial number of parameters in models. As a method of combining multiple levels of data, then, the dummy variables approach fails. Moreover, *even if* dummy variables yielded interesting information about contextual variation, precisely how individual-level data relate to the macro-context ostensibly "measured" through the dummy indicators is still unaccounted for.[20]

The Interactive Model. The problems raised in the previous two sections can be alleviated somewhat through what we will call the "interactive model." In the linear modeling context, the interactive model treats the relationship between an independent variable and a response variable as nonadditive and one that is mediated through one or more independent variables. With multiple levels of data, we might think the relationship between individual-level variables and the response variable as being mediated by some extra-individual variable measured at a higher level. For example, we could postulate that

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 z_{1j} + \beta_3 x_{1i} z_{1j} + \varepsilon_i ,$$

where z_{ij} denotes a unit specific or extra-individual (hence the *j* subscript) variable and β_3 denotes an interaction "effect" between the individual-level variable x_1 and $z_{1j.}$ Dropping subscripts for now, if one or both *z* and *x* are continuous or semi-continuous variables, then the interaction parameter, β_3 illustrates how the bivariate slope between *x* and *y* is mediated by or varies with values of *z*. That is, *x* "interacts" with *z* to produce varying slopes. Furthermore, if *z* is an indicator of a hypothesized contextual factor (for example, a nation's unemployment rate, type of government, ideological climate), then we have, apparently, modeled the contextual effect.

The interactive model, at least when compared to the "separate models" or dummy variables approach, is clearly more attractive in terms of combining multiple levels of data. Instead of ignoring contextual factors as is usually done with separate regressions and instead of collapsing the contextual "effect" into the form of a dummy variable, we have both accounted for unit-specific contextual variables, pooled the data across units, *and* linked the contextual factor to the individual-level variable. All of this yields parameter estimates allowing us to consider how individual observations "move" or vary with contextual variables.

Obviously, the interactive model is well known to political scientists (in large part due to Friedrich's [1982] article) and has been frequently used to demonstrate contextual effects, election effects, country effects, group effects and the like (c.f. Iverson 1991). The problem with the interactive model, however, is the implicit specification that the contextual "effect" is *deterministic*. To see this, we can retrieve the interactive model through the following exercise. Suppose we specify the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i ,$$

but believe that the relationship between x_{1i} and y_i is mediated by the contextual variable z_{1j}. Under such conditions, we can rewrite β_1 as

$$\beta_1 = \gamma_{10} + \gamma_{11} z_{1j}$$

and express β_0 as

$$\beta_0 = \gamma_{00} + \gamma_{01} z_{1j}$$

Substituting these expressions into the original model, we obtain

$$y_i = \gamma_{00} + \gamma_{01} z_{1j} + \gamma_{10} x_{1i} + \gamma_{11} z_{1j} x_{1i} + \varepsilon_i ,$$

which is equivalent to the original interactive model discussed earlier. But because in this formulation, the interaction of x and z is explicitly specified in terms of a varying coefficient, it is easy to see the deterministic nature of the interactive model. The two expressions for β_1 and β_0 are fully deterministic functions of z. This is equivalent to saying there is no stochastic disturbance associated with the interaction, and that the slope and intercepts are determined solely by z.

But suppose the interaction is *not* fully determined by z, but also is a function of stochastic error. Then the expressions for β_1 and β_0 can be written as

$$\beta_1 = \gamma_{10} + \gamma_{11} z_{1j} + \delta_{1j}$$
$$\beta_0 = \gamma_{00} + \gamma_{01} z_{1j} + \delta_{0j}$$

and substituting these into the interactive model produces

$$y_i = \gamma_{00} + \gamma_{01} z_{1j} + \gamma_{10} x_{1i} + \gamma_{11} z_{1j} x_{1i} + \delta_{0j} + \delta_{1j} x_{1i} + \varepsilon_i .$$

The two δ terms represent the stochastic disturbances associated with the slope and intercept. In this case, both the slope and intercept have been rewritten as *random coefficients*. The addition of the two error terms substantially complicates matters because instead of a single source of stochastic variance in the interactive model, we now have two sources. The first source emanates from the lower level data, the second from the contextual or higher-level data. Typically, this second source of stochastic variance is ignored in interactive models thus implicitly treating slopes and intercepts as stochastic functions of *z*. Consequently, the traditionally estimated interactive model is problematic in terms of its ability to combine multiple levels of data. Because the complex error structure that almost surely exists in many applied settings of comparative research is ignored through the standard interactive model, we turn our attention to multilevel methods.

To summarize this section, we have found that traditional methods used to combine multiple levels of data breakdown in important ways. While the standard interactive model ostensibly demonstrates contextual effects, we find it makes very strict assumptions about the relationship between slopes and intercepts and the contextual variables. More generally, the problem with the traditional methods is that they fail to capture "real" contextual factors (as in the case of the dummy variables approach), fail to relate higher-level data to lower-level data (as in the case of the separate models approach and dummy variables approach), and fail to account for macro-level stochastic variation across contextual units (as in the case of the standard interactive model, the dummy variables approach, and the separate models approach).

THE MULTILEVEL MODEL

In this section, we derive the basic form of the multilevel model and provide extensions. This section relies extensively on the work of Jackson (1991) and especially on the pioneering work of Bryk and Raudenbush (1993), Goldstein (1995), and Longford (1993). The intellectual roots of the multilevel model extend at least as far back as Swamy (1970), with his groundbreaking work on random coefficients models.

Basics of Multilevel Models

The simplest multilevel model that can be formulated considers only two levels of analysis. The first and most elementary of these levels will be referred to as level-1 and it is on this level that the analysis is focused. The remaining level is referred to as level-2 and provides the context for the level-1 units. For instance, level-1 units could be voters who are nested in different countries (level-2 units). The dependent variable is measured for level-1 units, since this is the primary level of analysis.[21] We shall denote the dependent variable as y_{ij}, where *i* refers to level-1 units and *j* refers to level-2 units. We assume there are J level-2 units, each containing n_j units.

The objective of multilevel models is to account for the expected value of y_{ij}. In the simplest case this is done via a linear model, although multilevel models may be nonlinearly specified. To simply further (again, without loss of generality) we first

consider only a single level-1 predictor, x_{ij}, in the level-1 model. The basis of the multilevel model is

[1] $$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij},$$

so that $E[y_{ij}] = \beta_{0j} + \beta_{1j} x_{ij}$, assuming $E[\varepsilon_{ij}] = 0$. The model in equation [1] looks very similar to a bivariate regression model, with one important exception: the regression parameters are subscripted in *j*. This indicates that, unlike normal regression analysis, the effect of level-1 is not considered fixed but allowed to vary across level-2 units. As we have argued, such variation is often assumed in comparative research, thus making the level-1 model attractive.

To model the contextual variation in regression parameters it is possible to formulate additional equations, this time for the contextual or level-2 units. One or both level-1 regression parameters constitute the left-hand side variables in these equations. The right-hand side variables consist at a minimum of a constant and typically also include at least one level-2 predictor and a disturbance. Thus, a typical level-2 model consists of the following equations:

[2a] $$\beta_{0j} = \gamma_{00} + \gamma_{01} z_j + \delta_{0j}$$

[2b] $$\beta_{1j} = \gamma_{10} + \gamma_{11} z_j + \delta_{1j}.$$

Here, z_j denotes a level-2 predictor, the parameters γ indicate fixed effects (similar to the coefficients in the classical linear regression model), and the parameters δ are disturbances that capture any random variation in the level-1 parameters that remains after controlling for the level-2 predictor.

The multilevel model is characterized by the complete system of equations that is given in equations [1]-[2b]. However, for the sake of simplicity the model is often characterized by a single equation by substituting [2a]-[2b] into [1]:[22]

[3] $$\begin{aligned} y_{ij} &= (\gamma_{00} + \gamma_{01} z_j + \delta_{0j}) + (\gamma_{10} + \gamma_{11} z_j + \delta_{1j}) x_{ij} + \varepsilon_{ij} \\ &= \gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} z_j x_{ij} + \delta_{0j} + \delta_{1j} x_{ij} + \varepsilon_{ij} \end{aligned}.$$

The first four terms on the right-hand side of equation [3] indicate the fixed effects in the model. The first of these terms gives the intercept or constant, the second the effect of the level-2 predictor, the third the effect of the level-1 predictor, and the fourth the interactive effect between the level-1 and level-2 predictors (the so-called cross-level interaction). The latter term provides insight into how contexts alter the impact of level-1 predictors. The last three terms in equation [3] are random effects – collectively they comprise the disturbance of the multilevel model. Here, δ_{0j} gives the residual contextual variance in the level-1 intercept after controlling for z_j, $\delta_{1j} x_{ij}$ gives the residual contextual variance in the slope for x_{ij}, and ε_{ij} is the usual level-1 disturbance term (capturing omitted level-1 predictors, measurement error in y_{ij}, and any idiosyncratic sources of variation in y_{ij} that can be attributed to level-1 units). We can conceive of δ_{0j} and $\delta_{1j} x_{ij}$ as *parameter noise*

and of ε_{ij} as *level-1 unit noise*. Thus, prediction errors of the multilevel model have two sources: (1) imperfect modeling of the dependent variable (ε_{ij}); and (2) imperfect modeling of the level-1 parameters (δ_{0j} and $\delta_{1j}x_{ij}$).

Equation [3] looks similar to the interactive contextual model that we discussed earlier. In fact, it should now be clear that the standard interactive model is a special case of the multilevel model in equation [3]: if we set $\delta_{0j} = 0$ and $\delta_{1j} = 0$, then equation [3] reduces to the interactive model. Notice, however, that this simplification depends on a rather strong assumption, namely that the contextual variation in intercept and slope can be perfectly accounted for by z_j (see equations [2a]-[2b]). In most cases this assumption is highly problematic because it assumes a far greater knowledge about contextual effects than we typically possess. Rather than assuming perfect predictability of contextual parameter variation, we should test for this; multilevel modeling allows one to do this.

By incorporating parameter noise terms, the multilevel model bypasses the dubious CLRM assumptions of homoskedasticity and no serial correlation. Indeed, it is easily verified that the variance of the multilevel disturbance term is not constant and that the disturbances from level-1 units within the same level-2 unit are correlated. To show this we can write the multilevel disturbance term as $u_{ij} = \delta_{0j} + \delta_{1j}x_{ij} + \varepsilon_{ij}$. Further, we shall make the following assumptions about the components of this disturbance:

A.1 $E[\delta_{0j}] = E[\delta_{1j}] = E[\varepsilon_{ij}] = 0$

A.2 $V[\delta_{0j}] = \tau_{00}, V[\delta_{1j}] = \tau_{11}, V[\varepsilon_{ij}] = \sigma^2$

A.3 $Cov[\delta_{0j}, \varepsilon_{ij}] = Cov[\delta_{1j}, \varepsilon_{ij}] = Cov[\varepsilon_{ij}, \varepsilon_{kl}] = 0$

A.4 $Cov[\delta_{0j}, \delta_{1j}] = \tau_{01}$

Assumption A.1 states that there is no systematic parameter noise or level-1 noise. Assumption A.2 states that parameter noise and level-1 noise can be characterized by constant variances.[23] Assumptions A.3 and A.4 indicate that the different components of u_{ij} are uncorrelated, with the exception of δ_{0j} and δ_{1j}. This means: (1) that there is no serial correlation between level-1 disturbances; and (2) that level-1 and level-2 disturbances are uncorrelated. The latter assumption implies that omitted level-1 predictors are not correlated with omitted level-2 predictors.

With these assumptions we can now derive the variance of u_{ij}:

$$\begin{aligned} V[u_{ij}] &= E\left[\left(\delta_{0j} + \delta_{1j}x_{ij} + \varepsilon_{ij}\right)^2\right] \\ &= E\left[\delta_{0j}^2\right] + 2x_{ij}E\left[\delta_{0j}\delta_{1j}\right] + x_{ij}^2 E\left[\delta_{1j}^2\right] + E\left[\varepsilon_{ij}^2\right] \\ &= \tau_{00} + 2x_{ij}\tau_{01} + x_{ij}^2\tau_{11} + \sigma^2. \end{aligned} \qquad [4a]$$

We see that while the components of u_{ij} are homoskedastic, u_{ij} itself is inherently heteroskedastic as it is a function of level-1 predictors. Indeed, only when no parameter noise is assumed for the level-1 slope, will $V[u_{ij}]$ be homoskedastic.

The multilevel disturbances are also serially correlated for level-1 units in the

same context. Let u_{ij} and u_{kj} denote two such disturbances, then:

$$
\begin{aligned}
Cov[u_{ij}, u_{kj}] &= E\left[\left(\delta_{0j} + \delta_{1j} x_{ij} + \varepsilon_{ij}\right)\left(\delta_{0j} + \delta_{1j} x_{kj} + \varepsilon_{kj}\right)\right] \\
&= E\left[\delta_{0j}^2\right] + x_{ij} E\left[\delta_{0j}\delta_{1j}\right] + x_{kj} E\left[\delta_{0j}\delta_{1j}\right] + x_{ij} x_{kj} E\left[\delta_{1j}^2\right] \\
&= \tau_{00} + \left(x_{ij} + x_{kj}\right)\tau_{01} + x_{ij} x_{kj} \tau_{11} \\
&\neq 0.
\end{aligned}
\qquad \text{[4b]}
$$

Clearly, when the level-1 parameters are modeled as stochastic functions of level-2 predictors, multilevel disturbances will be correlated for units in the same context. Only when we assume perfect predictability of the level-1 parameters can we safely assume that there is no serial correlation.

The General Multilevel Model and Sub-Models

We can generalize equations [1]-[3] by including multiple level-1 and level-2 predictors. Let there be P level-1 predictors, x_{pij} (p = 1, ..., P). Then, the level-1 model is given by:

$$y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} x_{pij} + \varepsilon_{ij} . \qquad \text{[5]}$$

Further, assume that there are Q level-2 predictors, z_{qj} (q = 1, ..., Q). Then, the level-2 model for the intercept is given by:

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + \delta_{0j} , \qquad \text{[6a]}$$

and the level-2 model for the slopes is given by:

$$\beta_{pj} = \gamma_{p0} + \sum_{q=1}^{Q} \gamma_{pq} z_{qj} + \delta_{pj} . \qquad \text{[6b]}$$

Substitution of equations [6a]-[6b] into equation [5] gives the general linear 2-level model:

$$y_{ij} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + \sum_{p=1}^{P} \gamma_{p0} x_{pij} + \sum_{q=1}^{Q}\sum_{p=1}^{P} \gamma_{pq} z_{qj} x_{pij} + \delta_{0j} + \sum_{p=1}^{P} \delta_{pj} x_{pij} + \varepsilon_{ij} . \qquad \text{[7]}$$

The meaning of the various components in this equation is identical to that in equation [3].

The general model contains a wide variety of sub-models that are well-known in political science. Table 1 lists these models with the components of equation [7] that are required to derive them. Although all of these models are familiar, we shall spend some time describing each.

Table 1:
The General Multilevel Model and Its Sub-Models

Model	Model Components Included		
	Parameter Noise	Level-1 Predictors	Level-2 Predictors
General Multilevel Model	Yes	Yes	Yes
Random Coefficients Model	Yes	Yes	No
Means-as-Outcomes Model	Yes	No	Yes
Random Effects ANOVA	Yes	No	No
Interactive Contextual Model	No	Yes	Yes
Fixed Effects ANOVA	No	No	No

(1) *Random Coefficients Model.* In the random coefficients model, which is widely used in the analysis of pooled cross-sections and time-series data (e.g., Dielman 1989; Stimson 1985), the level-2 predictors are dropped from equations [6a] and [6b]. Thus, the level-1 parameters are conceived of as simple functions of a constant effect and random noise. This conceptualization results in the following model:

$$y_{ij} = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} x_{pij} + \delta_{0j} + \sum_{p=1}^{P} \delta_{pj} x_{pij} + \varepsilon_{ij} .$$

The level-2 predictors and cross-level interactions disappear from the general multilevel model, but the disturbances remain heteroskedastic and serially correlated.

(2) *Means-as-Outcomes Model.* In this model no level-1 predictors are included, so that the level-1 model simply consists of an intercept that is allowed to vary contextually. However, the model for this intercept does include level-2 predictors. Consequently the model is given by:

$$y_{ij} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + \delta_{0j} + \varepsilon_{ij} ,$$

so that the mean of y_{ij} is considered to be the outcome of contextual factors: $E[y_{ij}] = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj}$. The disturbances for this model, $u_{ij} = \delta_{0j} + \varepsilon_{ij}$, are homoskedastic ($V[u_{ij}] = \tau_{00} + \sigma^2$) but serially correlated ($Cov[u_{ij}, u_{kj}] = \tau_{00} \neq 0$).

(3) *Random Effects ANOVA.* The means-as-outcomes model can be modified by dropping the level-2 predictors from the model. This results in the random effects ANOVA model:

$$y_{ij} = \gamma_{00} + \delta_{0j} + \varepsilon_{ij} ,$$

where γ_{00} is the grand mean of y_{ij}. Random effects ANOVA is useful when the treatment levels are not considered fixed but sampled from a "population" of treatments (see Maxwell and Delaney 1989).

(4) *Interactive Contextual Model.* If the disturbances are removed from the level-2 model equations, the interactive contextual model that we discussed earlier is obtained. As noted before, the implicit assumption of this model is that the level-1 disturbances (which are the only disturbances left) are homoskedastic and not serially correlated. These are the assumptions of classical linear regression analysis.

(5) *Fixed Effects ANOVA.* Fixed effects ANOVA can be thought of as a modification of random effects ANOVA. Rather than assuming that the treatments are sampled, they are thought of as fixed: in each imaginable iteration of an experiment the same levels would be chosen over and over again. This being the case, parameter noise – which would arise from treatment sampling fluctuations – can be ignored. Consequently, we obtain the following model:

$$y_{ij} = \gamma_{00} + \varepsilon_{ij},$$

where γ_{00} is the grand mean of y_{ij}.

Extensions: Higher-Order Multilevel Models

The multilevel model cannot just be expanded to incorporate multiple predictors, it can also be extended across more than two levels of analysis. The basic logic here is a straightforward extension of the 2-level model: parameters at each level of analysis are allowed to vary contextually over the next-higher level of analysis, with the parameters at the highest level of analysis considered as fixed.

We can illustrate this logic for a three-level model (for example, of voters nested in different time periods in different states). Let y_{ijk} denote the dependent variable, with the added subscript *k* referring to the level-3 units (e.g., states). Then the level-1 model can be written as:

$$y_{ijk} = \beta_{0jk} + \sum_{p=1}^{P} \beta_{pjk} x_{pijk} + \varepsilon_{ijk}.$$

The parameters β are allowed to vary contextually across the level-2 units (for example, they could be time-varying parameters if the level-2 units are time periods) according to:

$$\beta_{0jk} = \gamma_{00k} + \sum_{q=1}^{Q} \gamma_{0qk} z_{qjk} + \delta_{0jk}$$

$$\beta_{pjk} = \gamma_{p0k} + \sum_{q=1}^{Q} \gamma_{pqk} z_{qjk} + \delta_{pjk}.$$

Finally, the parameters γ are allowed to vary contextually across the level-3 units.

Assuming S level-3 predictors, w_{sk}, with fixed effects λ, this implies the following set of equations:

$$\gamma_{00k} = \lambda_{000} + \sum_{s=1}^{S} \lambda_{00s} w_{sk} + \nu_{00k}$$

$$\gamma_{0qk} = \lambda_{0q0} + \sum_{s=1}^{S} \lambda_{0qs} w_{sk} + \nu_{0qk}$$

$$\gamma_{p0k} = \lambda_{p00} + \sum_{s=1}^{S} \lambda_{p0s} w_{sk} + \nu_{p0k}$$

$$\gamma_{pqk} = \lambda_{pq0} + \sum_{s=1}^{S} \lambda_{pqs} w_{sk} + \nu_{pqk} .$$

The structure that these seven equations produce is highly complex. Among other things, it contains main effects for the level-1, level-2, and level-3 predictors, double cross-level interactions, and triple cross-level interactions. In addition, the 3-level model contains an exceedingly complex disturbance term. Needless to say, extensions of the multilevel model to an even greater number of levels produce still more complex structures.

With recent advances in computational power, most software packages will now permit the analysis of at least 3-level models, with some allowing as many as nine levels of analysis (although cross-level interactions are usually not permitted in this case). However, we caution against moving beyond 2-level models, for two reasons. First, as the number of levels increases ever greater demands are placed on the data for estimating the parameters, in particular the variance components. As we will see, in these cases it becomes critical that sufficiently large numbers of level-2 and level-3 units are available and in we doubt this will be the case in most political science data sets. Second, the interpretation of complex multi-level models is very tricky. It forces us to think about contextually determined effects, in which the contextual determination is itself contextually determined, and so on. Perhaps as a general statement about the world, a supposition of such contingent contingencies is true, but it hardly makes for parsimony and it may well cause bewilderment rather than insight into the linkages between different levels of analysis. Thus, our recommendation is to refrain from using more than two levels, unless one has a clear rationale for including more levels and strong expectations about the nature of the effects and their contingencies.

Extensions: Nonlinear Multilevel Models

Another way in which multilevel analysis can be extended is by dropping the linearity assumption that has characterized our discussion thus far. Recent developments in multilevel analysis now permit for nonlinear model specifications and this opens the door to modeling discrete responses, event duration data, and counts.

The simplest multilevel model for discrete responses is that for binary variables. The most common model for such variables is the multilevel logit model, which modifies the linear multilevel model by specifying a logit link function (Goldstein 1991, 1995). Thus, the outcome of interest is the proportion of cases, π_{ij}, that fall into category 1 of

the binary outcome measure and the multilevel model for this proportion can be written in terms of the log-odds ratio:

$$\ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)=\gamma_{00}+\sum_{q=1}^{Q}\gamma_{0q}z_{qj}+\sum_{p=1}^{P}\gamma_{p0}x_{pij}+\sum_{q=1}^{Q}\sum_{p=1}^{P}\gamma_{pq}z_{qj}x_{pij}+\delta_{0j}+\sum_{p=1}^{P}\delta_{pj}x_{pij}+\varepsilon_{ij}\,.$$

It is conventional in this specification to set $V[\varepsilon_{ij}]=\sigma^2=1$, as is done in conventional logit models. However, the multilevel disturbance term continues to be heteroskedastic and estimation of the logit parameters is conducted assuming such heteroskedasticity (as opposed to homoskedasticity, as is the case with estimating conventional logit models).[24]

Most multilevel software packages require the use of a logit link function in the analysis of binary data. However, there has been some movement toward the use of the probit link function. Thus, the program MIXOR (Hedeker and Gibbons 1996) allows the user to either choose the logit or probit link function. Future simulation studies are needed to determine which of these functions behaves best in the context of multilevel analysis.

Another development associated with the program MIXOR is to extend multilevel analysis to ordinal dependent variables. Here the user again has a choice between logit or probit link functions, both of which are available in MIXOR. The statistical theory behind ordered multilevel logit and probit models is set out in Hedeker and Gibbons (1994; also see Goldstein 1995).[25]

Models for counts are also estimable in multilevel analysis by specifying a multilevel Poisson regression model (see Goldstein 1991, 1995). The outcome of interest in this model is the expected number of level-1 units displaying a particular outcome, n_{ij}^{a}, where the superscript *a* refers to the outcome of interest. This number can be expressed as:

$$n_{ij}^{a}=n_j\pi_{ij}^{a},$$

where n_j denotes the number of level-1 units in the j^{th} level-2 unit and π_{ij}^{a} denotes the probability of outcome *a* in the level-1 units. Most multilevel software packages proceed by modeling this probability using a log link function. Thus, the multilevel Poisson regression model can be written as:

$$\ln\left(\pi_{ij}^{a}\right)=\gamma_{00}^{a}+\sum_{q=1}^{Q}\gamma_{0q}^{a}z_{qj}+\sum_{p=1}^{P}\gamma_{p0}^{a}x_{pij}+\sum_{q=1}^{Q}\sum_{p=1}^{P}\gamma_{pq}^{a}z_{qj}x_{pij}+\delta_{0j}^{a}+\sum_{p=1}^{P}\delta_{pj}^{a}x_{pij}+\varepsilon_{ij}^{a}\,.$$

Typically, the variance of the level-1 disturbance term is set to 1 (which is in keeping with the assumption that it follows a Poisson distribution [Golstein 1995]). The multilevel disturbance term, however, is again heteroskedastic (and serially correlated).

From event counts it is possible to move to event duration models. Goldstein (1995) provides a discussion of various forms of the multilevel event history model. The motivation for multilevel methods as applied to event histories is the idea that individuals

or durations may be nested within some higher level unit and therefore, trajectories, failure times, and the like, may be influenced by variables measured at the extra individual-level. Although use of multilevel event history methods as not been widespread in the social sciences, especially political science, the methodology would seem very appropriate for many research questions, and particularly amenable to modeling heterogeneity problems that are rampant in political event history data (see Box-Steffensmeier and Jones 1997).

STATISTICAL THEORY

In the previous section we established that multilevel models are conceptually distinctive. They are also statistically distinctive, however, finding their roots in a body of statistical theory that is not common to political methodology. In this section we shall describe this theory, paying attention to the fundamental principles of estimation and testing of multilevel models. Since this is still an evolving field in the statistical literature, we shall discuss both mainstream and alternative approaches in the ensuing discussion, although we should note that only the mainstream approaches are currently implemented in standard multilevel software.

Preliminaries

To simplify the discussion of estimation theory, it is convenient to re-express the general multilevel model of equations [5]-[7] in terms of matrices and vectors. To do so, we collect the responses of all level-1 units in the j^{th} level-2 unit in a $n_j \times 1$ vector $\mathbf{y}_j$. Similarly, the responses on the level-1 predictors are collected into the $n_j \times (P + 1)$ matrix $\mathbf{X}_j$, which also includes the level-1 constant, and the level-1 disturbances are collected in the $n_j \times 1$ vector $\boldsymbol{\varepsilon}_j$. Finally, the level-1 coefficients are collected in the $(P + 1) \times 1$ vector $\boldsymbol{\beta}_j$. This allows us to express the level-1 model as:

$$[8] \qquad \mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j .$$

Furthermore, we collect all level-2 predictors into the $(P + 1) \times (Q + 1)$ matrix $\mathbf{Z}_j$, which includes the level-2 constant, all level-2 disturbances into the $(P + 1) \times 1$ vector $\boldsymbol{\delta}_j$, and all level-2 coefficients into the $(Q + 1) \times 1$ vector $\boldsymbol{\gamma}$. This gives the following expression for the level-2 model:

$$[9] \qquad \boldsymbol{\beta}_j = \mathbf{Z}_j\boldsymbol{\gamma} + \boldsymbol{\delta}_j .$$

Substitution of equation [9] into equation [8] gives the expanded form of the multilevel model:

$$[10] \qquad \mathbf{y}_j = \mathbf{X}_j\mathbf{Z}_j\boldsymbol{\gamma} + \mathbf{X}_j\boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j .$$

To complete the multilevel model we restate assumptions A.1.-A.2.:

$$\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$
$$\boldsymbol{\delta}_j \sim N(\mathbf{0}, \mathbf{T}),$$

where **T** is the variance-covariance matrix of the level-2 disturbances.

Estimation Theory

The most common method for estimating multilevel models is maximum likelihood estimation (MLE), whereby the fixed effects, level-1 coefficients, and variance components are estimated simultaneously. However, conceptually it is easier when we think of the estimation of the fixed effects, level-1 coefficients, and variance components as separate steps. When we do so, it becomes apparent that the estimation of multilevel models entails a mixture of generalized least squares (GLS), empirical Bayes (EB), and MLE methods.

Variance Components. Estimation of the variance components is the most controversial aspect of the estimation theory of multilevel models. Many programs estimate the variances of the disturbances through *Full MLE* (FML). This entails minimization of the deviance of the data, where deviance is defined as –2 times the log-likelihood function (for details and formulae see Bryk and Raudenbush 1992; De Leeuw and Kreft 1986; Longford 1993). However, other programs utilize a modification of MLE that is known as *Restricted MLE* (REML; Harville 1977). This method does not minimize the deviance of the data but the deviance of the least squares residuals (see Bryk and Raudenbush 1992; De Leeuw and Liu 1993; Longford 1993).

Many researchers advocate the use of RMLE, especially in small samples. The reason is that FML, while consistent and asymptotically efficient, does not adjust for the number of fixed effects that are estimated and hence tends to be biased. REML makes this adjustment and is hence, at least theoretically, the better of the two estimators.[26] This superiority should be evident, in particular, with small samples of level-2 units. The FML variance components will tend to be underestimated by a factor $\frac{J-(Q+1)}{J}$ compared to the REML variance components. For small samples of level-2 units (J is small), this reduction can be substantial. However, in large samples of level-2 units the reduction is generally uninteresting (see Cressie and Lahiri 1993).

Whether the differences between FML and REML are truly as dramatic in practical applications of multilevel modeling as the advocates of REML have sometimes suggested remains to be seen. Simulated comparisons of the two methods have not generated a clear picture of the circumstances under which one or the other method is preferred (Swallow and Monahan 1984). Evidence that we have seen (e.g., Kreft, De Leeuw, and Van Der Leeden 1994) fails to show dramatic differences in the FML and REML variance components, but this may be peculiar to the data that were used. Given the considerable unclarity about the status of FML and REML,[27] we suggest that users will try to use both and compare results. Alternatively, an entirely different approach to estimation could be taken, a topic that we shall address below.

Fixed Effects. Estimation of the fixed effects vector $\boldsymbol{\gamma}$ can be based on equation [9]. One approach would be to employ ordinary least squares (OLS), so that:

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{Z}_j'\mathbf{Z}_j\right)^{-1}\mathbf{Z}_j'\boldsymbol{\beta}_j .$$

Of course, $\boldsymbol{\beta}_j$ is unknown, but it can be estimated from equation [8] using OLS:

$$\hat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\mathbf{y}_j .$$

This is the approach of traditional contextual analysis (Boyd and Iversen 1979; also see Hanushek 1974). An important drawback of this approach is, however, that it does not take into account that $\hat{\boldsymbol{\beta}}_j$ is usually estimated with different levels of precision in different groups, if only because the sample sizes in the different groups differ. As a consequence, $\hat{\boldsymbol{\gamma}}$ is not BLUE (best linear unbiased), except in the special case of a balanced design in which the sample sizes for all level-2 units are identical.

A better approach is to use a *precision-weighted estimator* that gives greatest weight to those estimates of $\boldsymbol{\beta}_j$ that are the most precise. This can be done via generalized least squares:

$$\tilde{\boldsymbol{\gamma}} = \left(\mathbf{Z}_j'\boldsymbol{\Delta}_j^{-1}\mathbf{Z}_j\right)^{-1}\mathbf{Z}_j'\boldsymbol{\Delta}_j^{-1}\boldsymbol{\beta}_j ,$$

where $\boldsymbol{\Delta}_j$ is the weight matrix. The only problem is now to define $\boldsymbol{\Delta}_j$. Since we want to precision-weight the estimates of $\boldsymbol{\beta}_j$, it makes sense to base $\boldsymbol{\Delta}_j$ on the dispersion matrix of $\hat{\boldsymbol{\beta}}_j$. It is easily demonstrated that this dispersion matrix takes the form

$$\boldsymbol{\Delta}_j = \mathbf{T} + \mathbf{V}_j = \mathbf{T} + \sigma^2\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1},$$

where $\mathbf{T}$ is the variance-covariance matrix for $\boldsymbol{\delta}_j$ and $\mathbf{V}_j$ is the normal OLS dispersion matrix.[28] We can think of $\boldsymbol{\Delta}_j$ as consisting of two parts. The first part, $\mathbf{T}$, gives parameter dispersion: random variance in the parameters. The second part, $\mathbf{V}_j$, gives the variance of the level-1 noise. This can be thought of as error dispersion, as it reflects the true lack of fit of the model (see Bryk and Raudenbush 1992). In practice, $\boldsymbol{\Delta}_j$ is of course estimated, using the FML or REML variance component estimates.

The GLS estimator has several desirable properties. First, it is unique and BLUE. Second, the estimator is responsive to the data. When a sub-group sample is small, for example, the estimate of σ^2 tends to increase because the degrees of freedom are smaller. Consequently, the data from the sub-group will not be weighted as much in determining the estimate of the fixed effect. Finally, although they will receive less weight, even sparse samples can contribute to the estimation of $\tilde{\boldsymbol{\gamma}}$. Thus, no information needs to be thrown away. Moreover, since the estimation of $\tilde{\boldsymbol{\gamma}}$ is based on information

from all sub-groups, it does not matter very much if the sub-groups are small as long as the total sample size is large enough.[29]

Level-1 Coefficients. The estimation of the variance components and fixed effects is sufficient for evaluating the complete multilevel model of equation [10]. However, often it is important to also obtain estimates for level-1 coefficients for specific sub-groups. Indeed, such estimates are invaluable for the interpretation of multilevel models (see below) and should be obtained routinely as part of the output from multilevel software.

From a statistical perspective, two different estimators of the level-1 coefficients are available. First, one could simply consider the data of the sub-group of interest and obtain OLS estimates based on only those data. This approach focuses on the level-1 units in a particular sub-group and uses equation [8] for that group. The resulting estimates are collected in $\hat{\boldsymbol{\beta}}_j$.

Alternatively, it is possible to use equation [9] to obtain estimates of the level-1 coefficients for a sub-group. This approach focuses on the level-2 units and utilizes the principle that $E[\boldsymbol{\beta}_j] = \boldsymbol{\beta}_j = E[\mathbf{Z}_j\boldsymbol{\gamma} + \boldsymbol{\delta}_j] = \mathbf{Z}_j\boldsymbol{\gamma}$, so that $\hat{\hat{\boldsymbol{\beta}}}_j = \mathbf{Z}_j\tilde{\boldsymbol{\gamma}}$. Thus, by taking the estimates of the fixed effect and the sub-group information for the level-2 predictors it is possible to construct estimates of the level-1 coefficients.

Under the usual assumption of correctly specified (level-1 and level-2) models, the two alternative estimators for the level-1 coefficients are both unbiased. However, they are generally not equally precise. This yields a useful criterion for combining the two estimators: we can take their weighted average, where the weight attached to an estimator is determined by its precision. The resulting estimator is an *empirical Bayes* (EB) *estimator* that, as we shall, see has several attractive properties.[30]

The weights used to obtain the EB estimates are given by:

$$\boldsymbol{\Lambda}_j = \mathbf{T}(\mathbf{T} + \mathbf{V}_j)^{-1},$$

which is the ratio of parameter dispersion over total dispersion (parameter dispersion plus error dispersion). Using these weights the EB estimator for the level-1 coefficients is:

$$\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Lambda}_j\hat{\boldsymbol{\beta}}_j + (\mathbf{I} - \boldsymbol{\Lambda}_j)\hat{\hat{\boldsymbol{\beta}}}_j.$$

When the error dispersion is zero, $\boldsymbol{\Lambda}_j = \mathbf{I}$, and $\tilde{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_j$. When the parameter dispersion is 0, $\boldsymbol{\Lambda}_j = \mathbf{0}$, and $\tilde{\boldsymbol{\beta}}_j = \hat{\hat{\boldsymbol{\beta}}}_j$. In other cases, $\tilde{\boldsymbol{\beta}}_j$ shrinks to the most precise estimator, hence the alternative term of *shrinkage estimation* to denote EB.

EB estimation of the level-1 coefficients has several desirable properties. First, it can be demonstrated that the EB estimator produces a smaller mean-squared error than other estimators (Carlin and Louis 1996; Lindley and Smith 1972). Second, the EB estimator allows reliable estimation of level-1 coefficients even in sub-groups that are very sparsely populated. This is because the estimator considers $\tilde{\boldsymbol{\gamma}}$ which, as we have seen, is based on information from all sub-groups. Thus, other sub-groups can help in the

estimation of level-1 coefficients for sparsely populated sub-groups. This is often particularly desirable in comparative research, because problems of micronumerosity are common in this field (Western 1997; Western and Jackman 1995).

Estimation in Practice. Although several attempts at non-iterative estimation have been made (De Leeuw and Kreft 1986), most multilevel programs utilize an iterative procedure whereby the variance components and fixed effects are continuously and simultaneously updated until convergence takes place. A variety of algorithms are in use, including EM (implemented in HLM), iterative generalized least squares (implemented in ML3), and Fisher scoring (implemented in VARCL). Of these algorithms, EM tends to be the slowest and also is least likely to reveal problems when the model is off (see De Leeuw and Kreft 1995; Kreft, De Leeuw and Van Der Leeden 1994).

Alternative Estimators. While GLS and EB are central to all currently available multilevel programs, these estimators have always had their detractors. The general criticism is that both estimators critically depend on estimates of the variance and covariance components of the model, but neither one acknowledges that there may be uncertainty over these components (Bryk and Raudenbush 1992). Put differently, the FML and REML (co)variance component estimates are incorporated into GLS and EB as point estimates with a prior probability of one attached to them. But the specification of such a prior seems overly optimistic, in particular when the number of level-2 units is small. After all, in these cases conventional statistical theory indicates that estimators may not be very stable and that a slightly different sample could have produced rather different estimates.

To incorporate uncertainty over (co)variance components, three approaches have been proposed. The first, is to simply adjust the GLS and EB estimators by incorporating estimates of the sampling variance of the (co)variance components (Morris 1983; Kackar and Harville 1984), possibly via bootstrapping (Laird and Louis 1987). To date, this approach has only been implemented for very simple multilevel models and there is considerable doubt that its implementation in more complex cases is straightforward (Seltzer, Wong and Bryk 1996).

The second approach is to use a fully Bayesian analysis in which priors are defined over all parameters, including the (co)variance components. However, this approach is very demanding and has been implemented only for simple models (Rubin 1981). Moreover, the use of a fully Bayesian approach introduces its own uncertainties. What prior distribution should be specified, for example, for the variance components? Our theories of contextually determined behavior may be too limited to provide much guidance to questions like this, making the use of the fully Bayesian approach difficult – more a future prospect, than a viable research strategy at the present.

The third approach is to use data augmentation in a Bayesian approach, i.e., to use a Gibbs sampler. In this approach, which belongs to the set of Markov Chain Monte Carlo algorithms, the joint prior distribution over the parameters is subdivided so that it is possible the sample conditionally from the posterior of one parameter, taking the other parameters as given. Repeated sampling gives the desired information to make inferences (see Goldstein 1995).

Applications of the Gibbs sampler in multilevel analysis have been quite

successful, although computationally demanding (Seltzer, Wong and Bryk 1996; Zeger and Karim 1991). We believe that this approach holds great promise in the future for multilevel modeling in general, and applications in political science specifically. Multilevel data structures in political science often contain few level-2 units, making uncertainty over (co)variance components a legitimate issue. However, at present we do not envision wide scale use of the Gibbs sampler in practical applications of multilevel models, since currently no software exists to implement the procedure and since it has some problems itself, such as the difficulties involved in assessing conversion. For this reason none of the applications in this paper are based on the Gibbs sampler.

Hypothesis Testing

An important part of multilevel modeling involves testing parameters and models to see which parts of the multilevel model are statistically important. Hypothesis tests for multilevel models are readily available, although there is some disagreement over the appropriate test statistics.

Testing Models. The use of multilevel analysis will almost surely involve the assessment of different models. Presently there is no method of assessing the fit of a model by itself – the way there is, for example, in covariance structure analysis. At best there are diagnostics, most importantly the deviance and related statistics such as Akaike's Information Criterion and Schwarz's Bayesian Information Criterion. For all of these diagnostics the general rule of thumb is that smaller values are better, although one can never be sure how small is good enough.

A test is available when one model is pitted against another model. To do so we assume that the smaller of these models is nested within the larger model. Let D_1 denote the deviance for the smaller model and D_2 the deviance of the larger model, and let *m* denote the difference in the number of estimated parameters (fixed effects and (co)variance components). Then,

$$D_1 - D_2 \sim \chi^2_m,$$

where χ^2_m denotes the χ^2-distribution with *m* degrees of freedom.

To implement this test procedure, it is useful to settle on a baseline model for comparison. In many contexts, a useful baseline would be a fixed effects model that only contains level-1 predictors. From there on one can move to a random coefficients model to assess whether there is significant parameter dispersion. If this is the case, the third model that can be fitted is one that includes level-2 predictors of the intercept and/or level-2 predictors for slopes. An application of this model testing sequence can be found in the Illustrations section of this paper.

Testing Individual Variance Components. Tests of individual variance components typically involve the null hypothesis H_0: $\tau_{pp} = 0$, where τ_{pp} denotes a particular diagonal element in the matrix **T**. There are at least three different recommendations for performing a test of this hypothesis. First, Goldstein (1995) suggests the use of the model comparison test. Here two models are estimated, one including the (co)variance

component and the other omitting it, and the difference in deviances is referred to a χ^2-distribution. The advantage of this test procedure is that it can be readily extended to a joint test of multiple (co)variance components.

A second approach is to take the ratio of the square root of the variance component and its estimated standard error and refer this to a student's t-distribution with J – Q – 1 degrees of freedom (see Longford 1993). This approach works well when a (co)variance component is large but is suspect when it is close to 0, in which case the symmetry of the student's t-distribution is usually inappropriate (Bryk and Raudenbush 1992).

The third approach is to obtain a χ^2-distributed test statistic for a (co)variance component. This can be done by taking the sum of squared residuals for a particular level-2 model and dividing this by the estimate variance of the variance component involved in this model. The resulting test statistic follows a χ^2-distribution with J – Q – 1 degrees of freedom. This approach works well even when variance components are close to 0 and has the advantage over the model comparison approach that it is not necessary to estimate multiple models. However, this approach is not available in all software packages, with many relying on the student's t-distribution instead.

Testing Individual Fixed Effects. The model comparison approach can also be used to test the significance of individual fixed effects, but this is conventionally not done. The typical approach is akin to tests of fixed effects in classical linear regression and involves evaluating the test statistic:

$$\frac{\tilde{\gamma}_{pq}}{\sqrt{\hat{V}[\tilde{\gamma}_{pq}]}}$$

for the null hypothesis H_0:$\gamma_{pq} = 0$. This test statistic is referred to a student's t-distribution with J – Q – 1 degrees of freedom.

Other Tests. It is also possible to perform significance tests on level-1 coefficients in particular sub-groups. We shall not discuss such tests (see Bryk and Raudenbush 1992) because we would like to discourage them. First, given that level-1 coefficients would be tested in many sub-groups, one should be very careful in interpreting significance levels. Indeed, with so many tests Bonferroni adjustments would almost surely be necessary. More serious, however, is the fact that the tests (even with Bonferroni adjustments) will be too liberal, unless there are many level-2 units (Bryk and Raudenbush 1992).

We believe that significance tests of level-1 parameters are often used as an interpretative device to determine in which sub-groups an effect "matters." However, if this is the objective, then much better methods are available. One of these is to simply graph the regression lines for different sub-groups and eyeball the results. This gives insight in the substantive significance of level-1 predictors in particular sub-groups, which is any way a better criterion for assessing if an effect "matters" than statistical significance.

Measures of Fit

Often researchers want to know how much variance they have explained. In regression analysis this question is typically answered by referring to the coefficient of determination. It is possible to do the same in multilevel analysis, although in this case not one but multiple coefficients of determination will be obtained.

A coefficient of determination can first be defined for the level-1 model. Here the objective is to assess the ratio of error variance over total variance. Longford (1993) suggests the following R^2-measure for this purpose:

$$R_1^2 = 1 - \frac{\hat{\sigma}_P^2}{\hat{\sigma}_0^2},$$

where $\hat{\sigma}_P^2$ is the least squares estimate of the residual level-1 variance for a model with P level-1 predictors and $\hat{\sigma}_0^2$ is the least squares estimate of the residual level-1 variance for a model without any level-1 predictors. If the level-1 predictors can perfectly account for the dependent variable, $\hat{\sigma}_P^2 = 0$, and $R_1^2 = 1$. If the level-1 predictors add nothing to the explanation of the dependent variable, then $\hat{\sigma}_P^2 = \hat{\sigma}_0^2$, and $R_1^2 = 0$.

A coefficient of determination can also be computed for each level-2 model. Here the relevant comparison is between the parameter variance estimate for a random coefficients model and the parameter variance estimate for a model that contains level-2 predictors. Bryk and Raudenbush (1992) suggest the following R^2-measure:

$$R_{2p}^2 = \frac{\hat{\tau}_{pp}(\text{RCM}) - \hat{\tau}_{pp}(\text{SIOM})}{\hat{\tau}_{pp}(\text{RCM})},$$

where $\hat{\tau}_{pp}(\text{RCM})$ denotes the estimated parameter variance for the random coefficient model (RCM), $\hat{\tau}_{pp}(\text{SIOM})$ denotes the estimated parameter variance for the slope or intercept as outcome model, and R_{2p}^2 denotes the coefficient determination for the level-2 model for the p^{th} level-1 coefficient.

As always, extreme care should be taken in the interpretation of the coefficient of determination, especially since the coefficients discussed here are unadjusted. We think of these coefficients as crude diagnostics of model performance. However, to truly assess the comparative fit of different models, it is best to rely on the test that we outlined earlier.

Model Specification

We want to conclude our discussion of the statistical theory of multilevel models by considering the issue of model specification. The analysis of multilevel models necessitates model specification choices that researchers do not ordinarily have to make, and it is important to point out the issues that are involved. A first issue concerns centering of the predictor variables, while a second issue concerns specification of the (co)variance components.

Centering of Predictors. In most statistical models predictors are included "as is," i.e., as they appear in the raw data. However, in multilevel models this often causes problems. There are two reasons for this. First, the use of raw data often causes ill-conditioning, especially in models with cross-level interactions (see Aiken and West 1991). Second, the interpretation of multilevel results often suffers when predictors are incorporated in raw form. For instance, the level-1 intercept cannot be easily interpreted when a zero score is not a feasible outcome in the sample for any of the level-1 predictors. A similar argument can be made for the level-2 intercept.

Centering, then, is critical to multilevel modeling.[31] Indeed, it is so central that several software packages (e.g., VARCL) *automatically* center the data before estimation. In other cases it is left to the researcher to center the data. In this case there are generally two strategies (other than not centering) for centering the level-1 predictors: centering with respect to the grand mean, or centering with respect to sub-group means. For the level-2 predictors there are only two choices: not centering or centering around the grand mean.

In a recent review of the topic, Kreft, De Leeuw and Aiken (1995) conclude that the question of how to center is primarily a theoretical one, for statistically speaking different centering methods tend to yield equivalent results. The central question is what theoretical interpretation one wants to give to the level-1 and level-2 intercepts. In the absence of centering, the level-1 intercept is the expected value of the dependent variable when all level-1 predictors are 0. Moreover, the level-2 intercepts give the expected values of the level-1 intercepts and slopes when all level-2 predictors are 0. For these interpretations to have any validity, the zero-scores on level-1 and level-2 predictors should occur in the sample. When the level-1 and level-2 predictors are centered around the grand mean, the level-1 intercept gives the expected value of the dependent variable for level-1 units whose score on the level-1 predictors is the average across all level-2 predictors. In this case, the level-2 intercepts give the expected values of the level-1 intercepts and slopes for cases whose level-2 predictor score is the average. Finally, if the level-1 predictors are centered around the group mean, then the level-1 intercept gives the expected value of the dependent variable assuming that a level-1 unit's scores on the predictors are the average in a particular group.

One way to conceptualize the different centering methods is to consider what they call attention to. Adopting the natural metric of the level-1 and level-2 predictors (without centering) calls attention to a zero-score on those predictors – this score stands out. This makes perfect sense, for example, when the 0-score refers to a control group to which we want to make comparisons. Centering around the grand mean calls attention to the typical values for the level-1 predictors, regardless of where these values occur in terms of level-2 units. This type of centering is often useful when the primary interest is in the level-1 units and one wishes to assess the impact of level-1 predictors against some baseline value of the predictor for those units. Finally, centering around the sub-group means calls attention to context. The primary interest is now in the typical value of level-1 predictors within specific contexts and effects of these predictors are assessed against this baseline. This strategy may be particularly useful if there is a great deal of between sub-group variance in the level-1 predictor means, so that it makes sense to adopt different reference points for effects in different sub-groups. On the other hand, the smaller the between sub-group variance in level-1 predictor means is, the less relevant the choice between

centering around the grand mean and centering around the sub-group means becomes.

Minimum Specifications of (Co)Variance Components. Specification of the elements of the matrix of variance and covariance components should be driven primarily by theoretical considerations. Every estimated variance component in this matrix implies that one assumes some stochastic variation in a level-1 coefficient, and every estimated covariance component implies that one assumes that stochastic fluctuation in one level-1 coefficient are systematically related to stochastic fluctuations in another level-1 coefficient.

In general, social scientists have generally stronger theoretical reasons to specify variance components than covariance components. In practice, this implies that researchers often do not include all possible covariance components, since some may not be theoretically meaningful or interpretable. We see no problems with this practice, which can cut down considerably on computation time, except under three circumstances.

First, it is recommended that a covariance between the level-1 slopes and intercepts is always included. This is important because it typically is the case that level-2 units with distinctive values on the intercept also show distinctive values on the slope. Second, for obvious reasons, covariance components should be specified for dummy predictors that capture categories in the same underlying categorical variable. Finally, when one level-1 predictor is derived from one or several other predictors, it is advisable to specify covariance components between their slopes. This is most relevant in cases in polynomial type models or models containing interactions between level-1 predictors.

APPLICATIONS

The Ideological Basis of Support for European Integration

Background. Studies of support for European integration typically come in two forms. The first consists of aggregate level data and focuses mostly on cross-national variations and time trends in the average level of support (Eichenberg and Dalton 1993). The second consists of individual level data and focuses on factors that may lead individual citizens to support or oppose the EU (Deflem and Pampel 1996, Janssen 1991). Studies that combine the different types of data are few and far between, and when they have been conducted the primary focus in the individual level data has been mostly on objective demographic factors (Gabel and Palmer 1995). There may be a simple reason for this: subjective individual-level factors have been notoriously poor predictors of EU support, exerting miniscule effects, and any attempt at including them in an analysis seem doomed from the outset.

The analysis of the role of political ideology (left-right self-placement) provides a case in point. Wessels (1995) concludes, for example, that the effect of ideology on EU support is very weak. While, Deflem and Pampel (1996) do not draw this conclusion themselves, the ideology effect that they report is among the weakest in their analysis.

The question is why ideology plays such a small role in determining EU support. One could conclude that the issue of European integration is simply not ideological in nature. This may well be true, but before accepting this conclusion we should examine at least one alternative explanation, namely contextual variation in the impact of ideology. It is possible that ideology plays an important role in some countries but not others and that

sometimes it exerts a positive effect and in other cases a negative effect. This is consistent with the existing evidence (Wessels 1995) and also rings true from casual observation of how political parties from the left and right have positioned themselves on the EU issue. For example, in the early 1990s ideological differences over EU support between parties in Greece were non-existing, so that we would expect no effect from left-right self-placement on EU support. In the same period, ideological differences over European integration in Denmark and Britain were profound. However, in Denmark the right favored integration, whereas in Britain it was the left (Labour). Thus we would expect opposite effects for left-right self-placement in both countries. When these contextual differences are ignored, as is so typically done in comparative studies of EU support, it should not come as a surprise that the overall effect of ideology is so small – the inconsistent effects of individual countries simply cancel each other out.

Multilevel models allow us to test this second possibility by testing whether the variance component that is associated with ideology is statistically significant. If it is, this is evidence for contextual variation. In this case, we may explore country-level factors that can account for the contextual variation. This allows one to determine whether the contextual variation is random or systematic, i.e., predictable on the basis of systematic differences between countries. This layered approach, whereby we consider ever more comprehensive models, will be illustrated in this example.

Models. We consider a simple model of EU support that is patterned after the work of Deflem and Pampel (1996). The dependent variable here is a dichotomous measure of EU support that is based on the following question in Eurobarometer 42.0: "Generally speaking, do you think that (your country's) membership of the European Union is a good thing, a bad thing, or neither good nor bad?" We coded the response "good thing" as 1 and the remaining responses as 0.

As predictors of this EU support measure we include age, gender, education, subjective class, and ideology as predictors. The basic model allowing for contextual variation, then, is:

$$\ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \beta_{1j}Age + \beta_{2j}Gender + \beta_{3j}Education + \beta_{4j}Class + \beta_{5j}Ideo\log y + \varepsilon_{ij}$$

where the subscript *j* denotes a particular country and the variance of ε_{ij} is fixed at 1.

We consider three special cases of this model. In the first one all contextual variation is removed, so that in our notation of the multilevel model

$$\beta_{0j} = \gamma_{00}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}, \beta_{3j} = \gamma_{30}, \beta_{4j} = \gamma_{40}, \beta_{5j} = \gamma_{50} \ .$$

This produces the a standard logit model. In our second model specification, all effects are considered fixed except for the intercept and the coefficient for ideology. For these two effects we stipulate the following equations:

$$\beta_{0j} = \gamma_{00} + \delta_{0j}$$
$$\beta_{5j} = \gamma_{50} + \delta_{5j}.$$

This results in a (partially) random coefficients model.

The final model we consider introduces country-level predictors for β_{0j} and β_{1j}. We consider three such predictors. First, we expect EU support among citizens to be greater in countries in which political parties are generally favorably disposed toward European integration. This follows an elite-driven model of public opinion (Wessels 1995). Second, we expect that EU support is less in countries in which the issue of European integration is highly salient among political parties, because this increase has typically been created in a climate of deep internal divisions over integration. One dimension of such divisions is ideology, so that we expect the effect of ideology to be stronger in countries in which the EU issue was salient. Finally, the difference in EU support between parties of the left and right may matter for average levels of EU support because it is one indicator of internal division. Moreover, this predictor should interact significantly with ideological self-placement. Specifically, in countries in which the left and right cannot be clearly distinguished in terms of EU support, we should expect no effect of ideological self-placement; in countries in which the left is clearly more pro-EU than the right, we should expect citizens from the left to be more supportive than those from the right; and in countries in which the right is most pro-EU, we should see the reversed pattern. Thus, our third model includes the following level-2 equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Party Support} + \gamma_{02}\text{Salience} + \gamma_{03}\text{Left} - \text{Right Difference} + \delta_{0j}$$
$$\beta_{5j} = \gamma_{50} + \gamma_{52}\text{Salience} + \gamma_{53}\text{Left} - \text{Right Difference} + \delta_{5j}$$

Data and Measures. The data concerning level-1 (individual-level)predictors come from Eurobarometer 42.0. Age is measured in years, education as the age at which the highest level of education was completed, gender is coded 0 for women and 1 for men, subjective class consists of five categories ranging from working class to upper class, and ideology is measured on a 10-point scale where 1 indicates the extreme left and 10 the extreme right.

The level-2 (country) data were collected by Ray (1997). Using an expert survey, Ray coded the support level for European integration as well as the salience of this issue for all parties in a country. The "Party Support" measure that we use is the average support of all parties weighted by their electoral representation. Salience is similarly defined. Finally, "Left-Right Difference" is a 3-category measure where –1 indicates that the left is clearly more favorable toward European integration than the right, 0 indicates that there are no clear differences between the left and right, and 1 indicates that the right is clearly more favorable toward integration than the left.

We consider data for 9540 citizens from 11 countries: Belgium, Denmark, France, Germany, Greece, Ireland, Italy, the Netherlands, Portugal, Spain, and the U.K. While the number of level-2 units is not very large, we shall see that it is still possible to run a multilevel analysis and obtain interesting results from it.

Table 2:
Different Models of Support for European Integration [a]

	Model 1		Model 2		Model 3	
Effect	**Estimate**	**est. s.e.**	**Estimate**	**est. s.e.**	**estimate**	**Est. s.e.**
Level-1 Main Effects:						
Constant	-.951*	.175	-1.204	[b]	8.948	[b]
Age	-.003*	.001	-.001	.001	-.002*	.001
Gender	.232*	.044	.225*	.044	.233*	.045
Education	.055*	.009	.072*	.009	.073*	.009
Class	.215*	.022	.186*	.023	.193*	.023
Ideology	.024*	.011	.015	.047	-.892*	.338
Level-2 Main Effects:						
Party Support					-.387	.369
Salience					-2.560*	.649
Left-Right Difference					-11.033	28.271
Cross-Level Interactions:						
Ideology × Salience					.291*	.108
Ideology × L-R Diff.					.080	.050
Variance Components:						
Constant			.271*	.113[c]	.152*	.086
Ideology			.023*	.034[c]	.013*	.027
Deviance	12096.243		11473.015		11430.742	

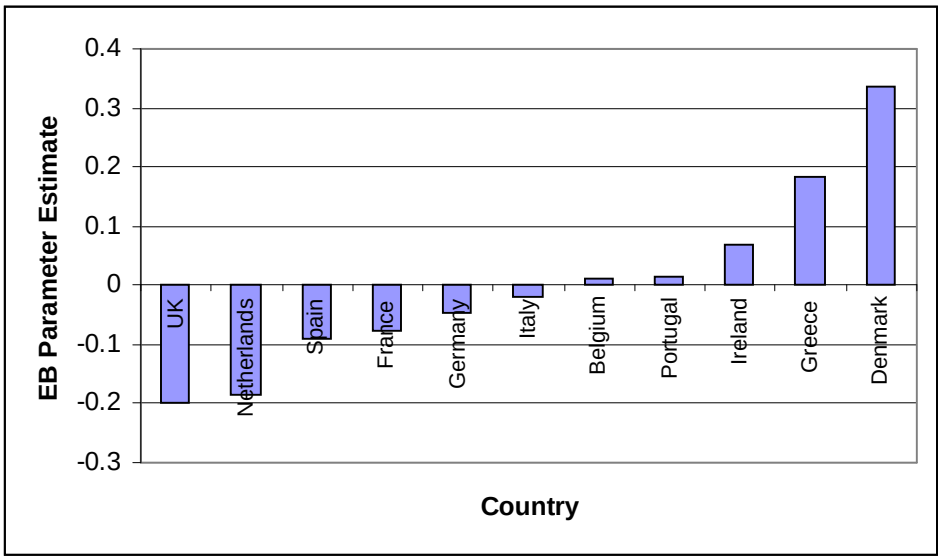
Notes: [a] Model 1-estimates were obtained in STATA using MLE; Model-2 and Model-3 estimates were obtained in VARCL using FML; [b] VARCL does not compute estimated standard errors for the constant; [c] estimated standard error is for the square root of the variance component.
* $p < .05$

Results. Table 2 gives the VARCL results for the three different models of EU support that we described. We shall first consider the relative fit of these models and will then comment on the parameter estimates. In terms of relative fit, Model 1 performs significantly worse than Model 2: The difference in deviance for these two models is 623.228, which at 3 degrees of freedom results in $p = .000$. Obviously, the poor fit of Model 1 relative to Model 2 is due in large part to the constraint that the constant is the same across all countries. However, this does not tell the entire story. When we compare Model 2 to a modified version of Model 1 with a random constant, we still observe a significant improvement in the fit of Model 2 (difference in deviance = 134.286, degrees of freedom = 2, $p = .000$). On the other hand, Model 2 clearly fits the data worse than Model 3: The difference in deviance is 42.273, which at 5 degrees of freedom gives $p = .000$.

The variance component for the ideology effect in Model 2 is not very large (.023), but clearly discernable from 0. We can use this variance component to obtain the

EB estimates of the ideology coefficients in different countries. These coefficients are depicted in the bar-chart in Figure 1. This figure lends support for our expectation that the impact of ideology on EU support runs the gamut from fairly large negative to fairly large positive effects. Fairly large negative effects are observed in the Netherlands and the U.K., whereas fairly large positive effects are found for Denmark and Greece. The latter two countries are also cases in which the salience of European integration to the national parties was high. Moreover, in Denmark the rightist parties were clearly more favorably disposed to European integration then the leftist parties, so that we should expect a positive effect for ideology. In the U.K. by contrast, where European integration was not very salient, it was the left that supported integration to a greater extent than the right, so that we should expect the negative ideology effect that we obtain. Thus, salience and left-right party differences seem to account for the patterns that we observe in Figure 1. The only anomaly here is the Netherlands, were we find right-leaning citizens to be less supportive than left-leaning citizens, even if the differences between parties go in the opposite direction. Moreover, salience of the EU issue was the lowest in this country, so that we would theoretically not expect any effect from ideology. We expect, then, that the EB estimates for Model 3 will continue to show the Netherlands as an anomalous case.

Figure 1:
Empirical Bayes (EB) Estimates of the Effect of Ideology in Different Nations



The estimates for Model 3 indicate that salience is the critical country-level predictor of EU support. Not only does it exert a significant (and as expected, negative) main-effect on the average support-level in a country, but it also interacts significantly with ideology. The nature of this interaction is such, that ideology indeed obtains a stronger effect the more salient the EU issue is. Moreover the effect changes from negative in low salience conditions to positive in high salience conditions, reflecting the

relative positions of Denmark and the U.K. None of the other level-2 predictors and cross-level interactions is statistically significant.

One question we should ask is whether the inclusion of salience (and the other level-2 predictors) is sufficient to eliminate the random variation in the constant and ideology effects. Inspection of the variance components indicates that this is not the case. While the variance components for these effects are about cut in half, they remain statistically significant. This may be an indication that Model 3 is under-specified, including too few level-2 predictors to account for the variation in constant and ideology effects. Because of the residual parameter variance for ideology, the EB estimate for anomalous case of the Netherlands remains sizable (-.155), as was expected.

Another question is what would happen if we were to ignore the residual random variation after the inclusion of the level-2 predictors and the cross-level interactions. Put differently, what problems would emerge if a contextual model were run that has the standard logit error term. On the whole, it turns out that the parameter estimates of a standard logit model are fairly close to those reported in Table 2 for Model 3. However, there is one major exception to this. In an ordinary logit analysis the effect of Left-Right Difference is estimated at only -.276, an indication, in our mind, of the problems that can arise in logit models when there is substantial heteroskedasticity (as the significant variance component for ideology implies). Moreover, the standard errors for the level-2 parameters and cross-level interactions are off in the standard logit model. As a consequence, one would reach different conclusions in that model than one would in the multilevel model. On statistical and theoretical grounds, the multilevel approach is preferable and should hence be the method of choice, as far as we are concerned.

CAVEATS OF MULTILEVEL MODELING

Although we have argued that multilevel methods are very well suited for many applications in comparative research, we would be remiss in not highlighting some "lowlights" of multilevel models. Less negatively, there are several caveats to consider before delving into multilevel modeling. Indeed, we have been careful to point out some of these pitfalls throughout and therefore, need not repeat ourselves here on issues of centering, interpreting levels of data greater than two, and issues of estimator selection. However, there are some conceptual issues worth discussing with regards to multilevel modeling.

Statistical Theory is Evolving. Although we discussed this earlier, this is one caveat that *does* bear repeating. For many of the models and estimators discussed in this paper, the statistical theory underlying the methodology is still in its infancy. And while a substantial body of statistics and econometrics has focused on random coefficients modeling (c.f. Swamy 1970, Hsiao 1986, Longford 1993), relatively little attention until recently, has been devoted to the issue of modeling multilevel data structures. Consequently, the properties of some of the estimators discussed herein are, quite frankly, not fully understood (Ita Kreft, personal communication). Yet because so many social theories, hypotheses, and data are hierarchically oriented, the "demand" and desire for these methods among applied researchers has substantially outweighed the "supply" of statistical theory. As a result, we caution that although one may be armed with

"contextual data", absent a *very strong* contextual theory, multilevel methods will be no savior.

Should We Be Concerned With Modeling "Context?" While contextual explanations of political behavior are widespread in political science, contextual analyses have been and remain controversial. At the heart of contextual modeling is an assumption that there is something interesting about how individuals are nested within aggregate "units" (or how level-1 units are nested within level-2 units). How contextual units are defined widely varies. The simplest demarcation of a contextual unit is probably spatial or geographic. Countries, states, *departements*, parishes, districts, and so forth, are easily definable, but are they politically important? As King (1996, 1997) has recently noted, the geographer's "modifiable areal unit problem" has substantial implications for political science, and in particular, contextual analysis. Roughly stated, the modifiable areal unit problem suggests that changes in the definition of the areal unit can, and generally does, elicit wild changes in interpretation of results (King 1997, 250-251). In terms of contextual analysis, arbitrary selection of contextual units or similarly, selection of units because of ease in data gathering can very likely produce misleading, or worse, irrelevant inferences. King (1996, 1997) persuasively notes that the modifiable areal unit problem is a *theoretical* problem and not an empirical problem. With regard to combining multiple levels of data, then, we stress that selection of the extra-individual unit *must* be theoretically driven, or else analyses of such data will suffer from the equivalent of the modifiable areal unit problem.

Blau (1980), in an essay on contexts and units in sociological research similarly articulates the theoretical problem of determining what the "right" unit is in contextual analyses. Blau notes that the unit forming the context (i.e. the "influencer") in some studies may, in other studies, be the object of the influence:

> In the sociological studies of social structures, the unit of analysis may range from small group to entire societies. Larger social structures encompass smaller ones, and the concepts and variables relevant for their investigation are not the same. Formal organizations can be the units of analysis in one investigation, but they may be the social context in another investigation of a narrower unit… . (Blau 1980 52).

An additional problem emerges when contextual units are treated as affiliational or associational groups. Contextual "effects" may be misleading because of self-selection bias (Hauser 1970; see also Hauser 1974). If individuals can select to which groups they are associated (which of course, they can), then research designs demonstrating a group-wise "social influence effect" may really be demonstrating nothing more than a self-selection mechanism. That is, the contextual effect is endogenous to the decision to join the group in the first place. And as Achen and Shively (1995) note, very little advancement has been made in solving this problem of contextual analysis.

But a more general criticism against contextual analysis has been leveled by King (1996) who argues that context "doesn't count" when it comes to explanations of individual-level political behavior. His argument, in part, centers on the premise that contextual effects are rarely robust in explaining behavior. Furthermore, political scientists should demonstrate that context does not "count" by theorizing and specifying models that are invariant across contextual units.

In general, we agree with each of these criticisms of contextual analysis. Demonstrating that individual-level outcomes vary across geographical units *without specifying the theoretical importance and significance* of this variation is tantamount to naming one's residuals. Spatial or "contextual" variation may be more-or-less a nuisance and inclusion of variables that indicate, for example, region, may help alleviate heterogeneity problems, but provide little substantive explanation of the political phenomenon. And calling such findings "contextual effects" doesn't improve inference making. Nevertheless, we contend that theories of contextual influence, at least in some quarters, extend well beyond the documentation and "discovery" of spatial variation. Huckfeldt and Sprague (1987, 1993, 1995) explicitly cast their work in terms of social interaction and not in terms of mere geographical variation. Additionally, Huckfeldt and Sprague, as well as other researchers (c.f. Przeworski 1974; Brown 1981, 1988; Noelle-Neumann 1984), have theorized that individual-level behavior *cannot* be understood apart from context, and any attempt to do so would elicit problematic inferences. This avenue of contextual analysis is largely derived from the work of social theorists like Durkheim, Bouden, Blau, and others who have theorized about the relationship between the individual and collectivities. The enterprise of contextual analysis, then, becomes an attempt to link aggregates and individuals theoretically and meaningfully, and not to solely demonstrate geographical variation. If one has no theory on how and why these levels of data relate, then contextual analysis devolves to the "naming your residuals" problem discussed previously. Additionally, if the aggregate-unit in which individuals are nested is arbitrarily or atheoretically selected, then too, will analyses of data fail to yield meaningful insights.

Measurement. As more complex statistical models are developed to combine multiple levels of data—the very models considered here—it seems clear to us that *greater* attention will have to be paid to issues of measurement theory, validity, and reliability assessment. How we define and measure concepts within the multilevel model is perhaps an even bigger issue than with traditionally less complex methods. Consider what is going on even in the simplest multilevel models. Lower level coefficients are treated as stochastic functions of variables created at a higher level. The variances and covariances within and between units are derived from level-1 coefficients and the values of these coefficients are "shrunk" to either the individual or the group-level. Level-2 coefficients are heavily predicated on measurement of level-2 attributes. But underlying these fireworks is a hefty premium on measurement and reliability. Bad measures in multilevel models "get worse" because such a heavy demand is placed on the data in terms of estimating level-1 and level-2 coefficients as well as the random parts. As comparativists move toward estimating these kind of models, we caution that substantial care needs to be taken in understanding how the measure reflects the theoretical "contextual variable" or gets at the appropriate individual-level variable.

CONCLUSION

In this paper, we have delineated the multilevel model in terms of comparative contextual analysis. Comparative analysis is replete with theories and hypotheses that posit a relationship between variables measured at multiple levels. Standard methodologies for combining multiple levels of data breakdown in important ways and

therefore provide an avenue toward multilevel modeling. Multilevel techniques provide leverage in linking multiple levels of data while at the same time avoid the pitfalls associated with traditional methods of dummy variable models, separate regressions, and standard interactive approaches.

REFERENCES

Achen, Christopher H. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Achen, Christopher H. 1983. "Toward Theories of Data: The State of Political Methodology." In *Political Science: The State of the Discipline*, ed. Ada F. Finifter. Washington, D.C.: American Political Science Association.

Achen, Christopher H. and W. Philips Shively. 1995. *Cross-Level Inference*. Chicago: University of Chicago Press.

Agnew, John. 1987. *Place and Politics: The Geographical Mediation of State and Society*. London: Allen and Unwin.

Agnew, John. 1996a. "Mapping Politics: How Context Counts in Electoral Geography." *Political Geography*. 15: 129-146.

Agnew, John. 1996b. "Maps and Models in Political Studies: A Reply to Comments." *Political Geography*. 15: 165-167.

Aiken, Leona S. , and Stephen G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage.

Alvarez, R. Michael and John Brehm. 1995. "American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values." *American Journal of Political Science.* 39: 1055-82.

Ames, Barry. 1987. "The Congressional Connection: The Structure of Politics and the Distribution of Public Expenditures in Brazil's Competitive Period." *Journal of Comparative Politics.* 147-172.

Ames, Barry. 1995. "Electoral Rules, Constituency Pressures and Pork Barrel: Bases of Voting in the Brazilian Congress." *Journal of Politics*. 57: 324-44.

Bartels, Larry. 1996. "Pooling Disparate Observations." *American Journal of Political Science.* 40: 905-42.

Basanez, Miguel, Ronald Inglehart, and Alejandro Moreno. 1997. *Human Beliefs and Values: A Cross-Cultural Sourcebook.* Ann Arbor: University of Michigan Press.

Beck, Nathaniel. 1983. "Time-varying Parameter Regression Models." *American Journal of Political Science*. 27: 557-600,

Beck, Nathaniel. 1985. "Estimating Dynamic Models Is Not Merely a Matter of Technique." *Political Methodology*. 11: 71-89.

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

Beck, Nathaniel and Jonathan N. Katz. 1996a. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis.* 6:

Beck, Nathaniel and Jonathan N. Katz. 1996b. "Lumpers and Splitters United: The Random Coefficients Model." Paper presented at the Annual Meetings of the Political Methodology Group, Ann Arbor, MI: July.

Beck, Nathaniel, Jonathan N. Katz, R. Michael Alvarez, Geoffrey Garrett, and Peter Lange. 1993. "Government Partisanship, Labor Organization, and Macroeconomic Performance: A Corrigendum." *American Political Science Review.* 945-948.

Blau, Peter M. 1980. "Contexts, Units, and Properties in Sociological Analysis." In *Sociological Theory and Research*, Hubert M. Blalock (ed.). New York: Free Press.

Box-Steffensmeier, Janet M., Laura W. Arnold, and Christopher J. W. Zorn. 1997. "The

Strategic Timing of Position Taking in Congress: A Study of the North American Free Trade Agreement." *American Political Science Review*. 91:324-338.
Box-Steffensmeier, Janet M. and Bradford S. Jones. 1997. "Time is of the Essence: Event history models in Political Science." *American Journal of Political Science*. (Forthcoming).
Boyd, L.H. , and G.R. Iversen. 1979. *Contextual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth.
Brown, Thad A. 1981. "On Contextual Change and Partisan Attitudes." *British Journal of Political Science.* 11: 427-48.
Brown, Thad A. 1988. *Migration and Politics*. Chapel Hill: University of North Carolina Press.
Bryk, Anthony S. , and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.
Bryk, Anthony, Stephen Raudenbush, and Richard Congdon. 1996. *Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs.* Chicago: SSI.
Cain, Bruce, John Ferejohn, and Morris P. Fiorina. 1987. *The Personal Vote: Constituency Service and Electoral Independence.* Cambridge: Harvard University Press.
Carlin, Bradley P. , and Thomas A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.
Collier, David. 1993. "The Comparative Method." In Ada W. Finifter (ed.) *Political Science: The State of the Discipline II*. Washington: American Political Science Association.
Cressie, N. , and S.N. Lahiri. 1993. "The Asymptotic Distribution of REML Estimators." *Journal of Multivariate Analysis* 45:217-233.
De Leeuw, Jan , and Ita G.G. Kreft. 1986. "Random Coefficient Models for Multilevel Analysis." *Journal of Educational Statistics* 11:57-85.
De Leeuw, Jan , and Ita G.G. Kreft. 1995. "Questioning Multilevel Models." *Journal of Educational and Behavioral Statistics* 20:171-189.
De Leeuw, Jan, and G. Liu. 1993. "Augmentation Algorithms for Mixed Model Analysis." Los Angeles: UCLA Department of Statistics. (Manuscript.)
Deflem, Mathieu , and Fred C. Pampel. 1996. "The Myth of Postnational Identity: Popular Support for European Unification." *Social Forces* 75:119-143.
Dielman, Terry E. 1989. *Pooled Cross-Sectional and Time Series Data Analysis*. 1. New York: Marcel Dekker.
Eichenberg, Richard C. , and Russell J. Dalton. 1993. "Europeans and the European Community: The Dynamics of Public Support for European Integration." *International Organization* 47:507-534.
Franklin, Mark, Michael Marsh, and Lauren McLaren. 1994. "Uncorking the Bottle: Popular Opposition to European Unification in the Wake of Maastricht." *Journal of Common Market Studies*. 32: 455-474.
Franklin, Mark and Wolfgang Rudig. 1995. "On the Durability of Green Politics: Evidence From the 1989 European Election Study." *Journal of Comparative Political Studies*. 3: 409-439.
Franklin, Mark, Cees Van Der Eijk and Michael Marsh. 1995. "Referendum Outcomes and Trust in Government: Public Support for Europe in the Wake of Maastricht."

Journal of West European Politics. 3: 101-116.
Friedrich, Robert J. 1982. "In Defense of Multiplicative Terms in Multiple Regression Equations." *American Journal of Political Science.* 26: 797-833.
Gabel, Matthew , and Harvey D. Palmer. 1995. "Understanding Variation in Public Support for European Integration." *European Journal of Political Research* 27:3-19.
Garrett, Geoffrey and Peter Lange. 1989. "Government Partisanship and Economic Performance: When and How Does `Who Governs' Matter?" *Journal of Politics*. 51: 676-93.
Geddes, Barbara. 1991. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis*. 2:131-50.
Geertz, Clifford. 1973. "Thick Description: Toward an Interpretive Theory of Culture." In Clifford Geertz (ed.) *The Interpretation of Cultures.* New York: Basic Books.
Gibson, James L. 1996. "A Mile Wide But an Inch Deep(?): The Structure of Democratic Commitments in the Former U.S.S.R." *American Journal of Political Science*. 396-420.
Gibson, James L. and Gregory A. Caldeira. 1996. "The Legal Cultures of Europe." *Law and Society Review*. 30: 55-85.
Gibson, James L. and Raymond M. Duch. 1992. "Anti-Semitic Attitudes of the Mass Public: Estimates and Explanations Based on a Survey of the Moscow Oblast." *Public Opinion Quarterly*. 56: 1-29.
Gibson, James L., Kent K. Tedin, and Raymond M. Duch. 1992. "Democratic Values and the Transformation of the Soviet Union." *The Journal of Politics*. 54: 329-72.
Goldstein, Harvey. 1991. "Nonlinear Multilevel Models with an Application to Discrete Response Data." *Biometrika* 78:45-51.
Goldstein, Harvey. 1995. *Multilevel Statistical Models*. 2. London: Edward Arnold.
Greene, William H. 1993. *Econometric Analysis* 2 ed. New York: Macmillan.
Haller, H. Brandon and Helmut Norpoth. 1994. "Let the Good Times Roll: The Economic Expectations of U.S. Voters." *American Journal of Political Science.* 38: 625-650.
Hanushek, Eric. 1974. "Efficient Estimates for Regressing Regression Coefficients." *The American Statistician* 28:66-67.
Hanushek, Eric A. and John E. Jackson. 1977. *Statistical Methods for Social Scientists*. San Diego: Academic Press.
Hardy, Melissa A. 1993. *Regression With Dummy Variables.* Newbury Park: Sage.
Harville, D.A. 1977. "Maximum Likelihood Approaches to Variance Component Estimation and Related Problems." *Journal of the American Statistical Association* 72: 320-340.
Hauser, Robert M. 1970. "Context and Convex: A Cautionary Tale." *American Journal of Sociology.* 75: 645-54.
Hauser, Robert M. 1974. "Contextual Analysis Revisited. *Sociological Methods and Research.* 2: 365-75.
Hedeker, Donald , and Robert D. Gibbons. 1994. "A Random-Effects Ordinal Regression Model for Multilevel Analysis." *Biometrics* 50:933-944.
Hedeker, Donald , and Robert D. Gibbons. 1996. "MIXOR: A Computer Program for Mixed-Effects Ordinal Probit and Logistic Regression Analysis." *Computer Methods and Programs in Biomedicine* 49:157-176.
Hibbing, John R. 1991. *Congressional Careers: Contours of Life in the U.S. House of*

Representatives. Chapel Hill: University of North Carolina Press.
Hogg, Robert V. and Allen T. Craig. 1978. *Introduction to Mathematical Statistics*. New York: Macmillan.
Hsiao, Cheng. 1986. *Analysis of Panel Data.* Cambridge: Cambridge University Press.
Huckfeldt, Robert and John Sprague. 1987. "Networks in Context: The Social Flow of Political Information." *American Political Science Review*. 81: 1197-1216.
Huckfeldt , Robert and John Sprague. 1993. "Citizens, Contexts, and Politics." In *Political Science: The State of the Discipline II*, Ada F. Finifter (ed.) Washington, D.C.: American Political Science Association.
Inglehart, Ronald. 1977. *The Silent Revolution: Changing Values and Political Styles Among Western Publics.* Princeton: Princeton University Press.
Iverson, Gudmund R. 1991. *Contextual Analysis*. Newbury Park: Sage.
Jackman, Robert W. 1985. "Cross-National Statistical Research and the Study of Comparative Politics." *American Journal of Political Science.* 29: 161-82.
Jackson, John E. 1992. "Estimation of Models with Variable Coefficients." *Political Analysis*. 3: 27-49.
Jackson, John E. and David C. King. 1989. "Public Goods, Private Interests, and Representation." *American Political Science Review.* 83: 1143-64.
Janssen, Joseph I.H. 1991. "Postmaterialism, Cognitive Mobilization and Public Support for European Integration." *British Journal of Political Science* 21:443-468.
Kackar, R. , and D. Harville. 1984. "Approximations of Standard Errors of Estimation of Fixed and Random Effects." *Journal of the American Statistical Association* 79:583-562.
Kahn, Kim Fridkin and Patrick J. Kenney. 1997. "A Model of Candidate Evaluation in Senate Elections: The Impact of Campaign Intensity." *Journal of Politics*. (Forthcoming).
Kalleberg, Arthur L. 1966. "The Logic of Comparison: A Methodological Note on the Comparative Study of Political Systems." *World Politics*. 19: 69-82.
Katz, Jonathan N. and Brian R. Sala. 1996. "Careerism, Committee Assignments, and the Electoral Connection." *American Political Science Review.* 90: 21-33.
King, Gary. 1990. "On Political Methodology." *Political Analysis*. 2: 1-29.
King, Gary. 1996. "Why Context Should Not Count." *Political Geography.* 15: 159-164.
King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior From Aggregate Data.* Princeton: Princeton University Press.
King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton: Princeton University Press.
Kingdon, John W. 1992. *Congressmen's Voting Decisions* 3rd Ed. Ann Arbor: University of Michigan Press.
Kramer, Gerald H. 1983. "The Ecological Fallacy Revisited: Aggregate Versus Individual-Level Findings on Economics and Elections and Sociotropic Voting." *American Political Science Reviews.* 77: 92-111.
Kreft, Ita G.G.,Jan De Leeuw , and Leona S. Aiken. 1995. "The Effect of Different Forms of Centering in Hierarchical Linear Models." *Multivariate Behavioral Research* 30:1-21.
Kreft, Ita G.G.,Jan De Leeuw , and Rien Van Der Leeden. 1994. "Review of Five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, ML3, and VARCL."

The American Statistician 48:324-335.
Laird, N. , and T. Louis. 1987. "Bootstrapping Empirical Bayes Estimates to Account for Sampling Variation." *Journal of the American Statistical Association* 82:739-757.
Lange, Peter and Geoffrey Garrett. 1985. "The Politics of Growth: Strategic Interaction and Economic Performance in Advanced Industrial Democracies, 1974-1980." *Journal of Politics*. 47: 792-827.
Lijphart, Arend. 1971. "Comparative Politics and Comparative Method." *American Political Science Reviews*. 65: 682-93.
Lindley, D.V. , and A.F.M. Smith. 1972. "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society, Series B* 34:1-41.
Longford, Nicholas T. 1993. *Random Coefficient Models*. 1. Oxford (U.K.): Clarendon Press.
MacKuen, Michael and Courtney Brown. 1987. "Political Context and Attitude Change." *American Political Science Review*. 81: 471-90.
Mason, William M., George Y. Wong, and Barbara Entwistle. 1983. "Contextual Analysis through the Multilevel Linear Model." *Sociological Methodology 1983-1984.*.
Maxwell, Scott E. , and Harold D. Delaney. 1990. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. 1. Belmont, CA: Wadsworth Publishing Co.
Mishler, William and Richard Rose. 1997. "Trust, Distrust, and Skepticism: Popular Evaluations of Civil and political Institutions in Post-Communist Societies." *Journal of Politics*. 59:418-451.
Morris, C.N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78:47-65.
Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry.* New York: Wiley-Interscience.
Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies.* Berkeley: University of California Press.
Ray, Leonard. 1997. "Measuring Party Orientations Towards European Integration: Results from an Expert Survey." (Manuscript).
Rivers, Douglas. 1988. "Heterogeneity in Models of Electoral Choice." *American Journal of Political Science*. 32:737-57.
Rokkan, Stein. 1966. "Comparative Cross-National Research: The Context of Current Efforts." In *Competing Nations*, ed. Richard Merritt and Stein Rokkan, 3-26. New Haven: Yale University Press.
Rubin, D. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics* 6:377-400.
Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review*. 64: 1033-53.
Sartori, Giovanni. 1991. "Comparing and Miscomparing." *Journal of Theoretical Politics.* 3: 243-57.
Sayrs, Lois W. 1987. *Pooled Time Series Analysis.* Newbury Park: Sage.
Seltzer, Michael H.,Wing Hung Wong , and Anthony S. Bryk. 1996. "Bayesian Analysis in Applications of Hierarchical Models: Issues and Methods." *Journal of Educational and Behavioral Statistics* 21:131-167.
Shively, W. Phillips. 1969. "`Ecological ' Inference: The Use of Aggregate Data to Study

Individuals." *American Political Science Review*. 63: 1183-96.
Shively, W. Phillips. 1974. "Utilizing External Evidence in Cross-Level Inference." *Political Methodology*. 1:61-74.
Shively, W. Phillips. 1987. "A Strategy for Cross-Level Inference Under an Assumption of Breakage Effects." *Political Methodology*. 11: 167-79.
Sprague, John. 1976. "Estimating a Bouden Type Contextual Model: Some Practical and Theoretical Problems of Measurement." *Political Methodology*. 3: 333-53.
Sprague, John. 1982. "Is There a Micro Theory Consistent with Contextual Analysis?" In *Strategies of Political Inquiry*. ed. Elinor Ostrom. Beverly Hills: Sage.
Stipak, Brian and Carl Hensler. 1982. "Statistical Inference in Contextual Analysis." *American Journal of Political Science*. 26: 151-75.
Stimson, James A. 1985. "Regression in Space and Time: A Statistical Essay." *American Journal of Political Science*. 29: 914-47.
Swallow, W.H. , and J.F. Monahan. 1984. "Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components." *Technometrics* 26:47-57.
Swamy, P.A.V. B. 1970. "Efficient Inference in a Random Coefficient Regression Model." *Econometrica*. 38: 311-323.
Wessels, Bernhard. 1995. "Development of Support: Diffusion or Demographic Replacement?" In *Public Opinion and Internationalized Governance*, ed. Oskar Niedermayer and Richard Sinnott. Oxford, UK: Oxford University Press.
Wessels, Bernhard. 1995. "Support for Integration: Élite or Mass-Driven?" In *Public Opinion and Internationalized Governance*, ed. Oskar Niedermayer and Richard Sinnott. Oxford, UK: Oxford University Press.
Western, Bruce. 1997. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." Typescript.
Western, Bruce , and Simon Jackman. 1995. "Bayesian Inference for Comparative Research." *American Political Science Review* 88:412-423.
Westlye, Mark C. 1991. *Senate Elections and Campaign Intensity*. Baltimore: Johns Hopkins University Press.
Woodhouse, G. 1995. *Multilevel Modelling Applications*. London: Institute of Education.
Zeger, S. , and M. Karim. 1991. "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach." *Journal of the American Statistical Association* 86:

NOTES

[1] Although it should be noted at the start that political methodologists have wrestled with problems of drawing inferences from multiple levels of data for some time (see Shively 1969, 1974, 1987, Sprague 1976, 1982, Achen 1983, Achen and Shively 1995, King 1990, 1996, 1997, just to name a few). In particularly, recent work by Achen and Shively (1995) and King (1997) has made significant inroads into the ecological inference problem. In this paper, we are concerned with the problem of *combining* individual-level and aggregate-level data, and of course therefore are assuming the researcher actually possesses individual-level data. Jackson (1992) dealt with this problem in his paper on variable coefficients models and we rely heavily on the ideas initially forwarded by him. Additionally, recent work by Beck and Katz (1995, 1996a, 1996b) and especially Western (1997), have examined the properties and application issues of random coefficients models for comparative political analysis.

[2] Although we are contrasting "individual-level" data with "aggregate-level" data, one need not perform analyses at the individual-level to use multilevel models. Multilevel approaches are generally applicable when one has data hierarchically nested (and one has a theory on how the multiple levels of data are related!). Thus, Western's (1997) work on multilevel models treats institutional attributes of countries as the lower-order unit and models unemployment rates across time.

[3] In contrasting "traditional" thick description methods with quantitative approaches, we do not intend to fan the fires of this debate in cross-area analysis. We think qualitative methods are an invaluable component to cross-area analysis and, if suitably rigorous in design (see King, Keohane, and Verba 1994), can yield inferences in many cases far stronger than quantitative models.

[4] King (1996) has recently called into the question the importance of actually modeling context. Later in the paper, we address King's argument. For now, we are assuming that "contextual variation" is a concern to cross-area analysts (although as we argue later, one need not model "context" to appropriately use multilevel models).

[5] Although important quantitative analyses of "small-n" cross-area data *have* been produced. See, for example, Lange and Garrett (1985), Garrett and Lange (1989) and Beck, Katz, Alvarez, Garrett, and Lange (1993).

[6] Although we suspect that in some quarters of cross-area research, King's (1997) solution will be viewed skeptically by those adhering to the view that aggregates posses "emergent properties." Because of these properties, aggregates (for example collectivities of individuals) possess a "reality" of their own making it impossible to decompose them into individual-level inferences. For example, Agnew (1996b) refers to King's ecological inference work as "ontological (and methodological) individualism" (165, parentheses in original). We generally agree with Achen and Shively's (1995) assessment of this view. They argue that "emergent properties" arguments, which are derived from social theorists like Durkheim , very often induce fallacious reasoning about aggregate data. As Achen and Shively note "[t]he Durkheimian beauties of emergent properties have often bedazzled researchers. Too much of the sociological literature on contextual effects has consisted of singing the theoretical praises of holistic effects, arguing the substantive plausibility of contextual effects, and then showing statistical biases due to aggregation effects, without noticing that meanings have shifted along the way" (Achen and Shively 1995 221).

[7] Clearly, individual-level comparative political data have been available for some time. Inglehart (1977) used individual-level data from the early Seventies to develop his theory of post-materialist values. The "development" we speak of really centers on the emergence of a diverse set of individual-level data collected across many global regions.

[8] And of course, in the United States, extensive individual-level and aggregate-level data have been available for decades from a variety of sources.

[9] Jackson (1992) also uses the example of legislative voting as a motivation for random coefficients models.

[10] Of course Hauser's (1970) admonitions are extremely relevant here. Hauser notes that group membership is largely self-selective. Thus, so-called "contextual effects" of group affiliation may reflect nothing more than selection biases, and *not* some group-level dynamic. We discuss in more detail at the end of the paper, some of Hauser's concerns with contextual analysis.

[11] Sociologist Peter Blau (1977, 1989) has been instrumental in theorizing about these kinds of social networks. Blau has argued that group or social identification is largely a function of demographic and socioeconomic factors such that individuals tend to identify more with individuals who possess similar attributes. Thus, geographical or familial connections are less important in terms of social networks than

demographic-based and socioeconomic-based ties. This kind of linkage or social context has prompted the concept of "Blau Space" (McPherson and Ranger-Moore 1991).

[12] Indeed, Mason, Wong, and Entwistle's (1983) early work on multilevel analysis was explicitly cast in terms of models for contextual analysis.

[13] In this context, the Chow test is simply an *F* test.

[14] Furthermore, it is not clear how useful this kind of testing is at all! Bartels' (1996) important work on the issue of pooling disparate observations suggests that *F*-tests of this sort frequently fail to test what "analysts need done" (Bartels 1996, 935). He notes that this test only focuses on goodness-of-fit and not on differences in parameter values.

[15] Beck (1985) in regard to time series data, and Rivers (1988) in regard to cross-sectional data have been instrumental in pointing out the problem associated with heterogeneity within subsets of data and proposed methods to address this problem. Subsequent work by Jackson (1991), Beck and Katz (1995a, 1995b), Bartels (1996), and Western (1997), among many others, have proposed methods to deal with these types of problems, but, as Bartels (1996) notes, it is commonplace for researchers to ignore problems associated with disparate observations.

[16] Greene (1993), we should note, makes this assertion in terms of his discussion of interaction terms with dummy variables. Using our example, separate regression estimates are identical to a pooled model with interaction terms between the covariates and a dummy variable denoting the country. And in fact if the disturbances across the country are equal, then it is most efficient to pool the observations rather than estimating separate regressions. Greene's (1993) point is that if the country-wise disturbance variances differ across groups (countries), then this dummy variable approach will not be feasible and it becomes most efficient to disaggregate the data.

[17] The "space" and "time" terminology of course stems from Stimson's (1985) classic article on pooled time series analysis.

[18] Though this approach is not always feasible. Inclusion of separate dummy variables for cross-sectional units, for example, may result in many hundreds of parameters.

[19] Or as Hanushek and Jackson (1977) note, "[t]he use of dummy variables admits to a lack of knowledge and/or data. We do not know the underlying cause of the differences in the populations, or we cannot break out separate elements of this different behavior" (103).

[20] Lest we sound too pessimistic about dummy variables approaches, we stress that our argument hinges on the assumption that the research is interested in making inferences from multiple levels of data., and not intent on soley alleviating problems of heteroskedasticity or autocorrelation.

[21] This aspect sets multilevel models apart from ecological models, in which the dependent variable (and all other variables) is measured at the aggregate level, for example for level-2 units.

[22] Certain software packages like SAS require the reformulation of the multilevel model into a single equation. Other packages (such as MLN [Goldstein 1995, Woodhouse 1996] and HLM4 [Bryk, Raudenbush, and Congdon 1996]) require specification of multiple equations.

[23] It is possible to extend the multilevel model by allowing for heteroskedastic level-1 disturbances (which may also be serially correlated – see assumption A.3). We will not discuss this extension in this paper but for an excellent discussion of the topic the reader is referred to Goldstein (1995).

[24] While the multilevel logit model is heteroskedastic, it is distinct from the heteroskedastic logit model described by Alvarez and Brehm (1996). In the terms of multilevel modeling, the source of heteroskedasticity in the heteroskedastic logit model is located in the level-1 units; the source of heteroskedasticity in the multilevel logit model is located in the level-2 units. Steenbergen is presently working on establishing a heteroskedastic multilevel logit model which combines the logic of the heteroskedastic logit model and that of multilevel logit analysis.

[25] Goldstein (1995) also discusses multilevel models for non-ordered polytomous variables. While in principle such models could be specified, we have not seen any applications of them, nor are we aware of software that will handle these models. The reason for this may be quite simple. Even in single-level analyses, multinomial logit and probit models can cause an enormous expansion of the number of parameters that needs to be estimated and this expansion will even be more extreme in multilevel specifications of these models.

[26] Although the statistical issues that are involved are much more complicated, we can compare the problem of FMLE to the one that emerges in estimating the sample variance. The variance MLE is given by

$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$, which differs from the common variance estimator $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$. The latter estimator takes into consideration that one degree of freedom is lost in estimating the sample mean that goes into the variance estimation. By making this adjustment, the common variance estimator eliminates the bias of the MLE (e.g., Hogg and Craig 1978).

[27] There is some evidence that RMLE may be badly behaved under certain circumstances. One of our colleagues noticed that the deviance for RMLE increased as he added more predictors to the model (George Rabinowitz, personal communication). This may have been a consequence of the size of the problem, which entailed a large number of predictors, but it may also reflect problems associated with the minimization of least squares residuals as opposed to the data.

[28] A complete proof can be found in Bryk and Raudenbush (1992). It is based on the well-known result that $\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j + \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\boldsymbol{\varepsilon}_j$. Substitution into equation [9], gives $\hat{\boldsymbol{\beta}}_j = \mathbf{Z}_j\boldsymbol{\gamma} + \boldsymbol{\delta}_j + \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\boldsymbol{\varepsilon}_j$. The dispersion matrix for $\hat{\boldsymbol{\beta}}_j$ is given by $V[\boldsymbol{\delta}_j] + V\left[\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\boldsymbol{\varepsilon}_j\right]$. An evaluation of this matrix gives the dispersion matrix as listed in the text.

[29] Of course, one may not want to include extremely sparse sub-groups (e.g., fewer than 5 cases), as the information in such sub-groups can often not be trusted.

[30] Empirical Bayes methods have in common that they evaluate Bayes' rule on the basis of the observed data (hence the name "empirical Bayes"). Such methods are now finding increasing acceptance in statistics, as they allow researchers to engage multiple estimators of a parameter in a consistent manner that has desirable statistical properties. For a discussion of EB methods, the reader is referred to the seminal work of Lindley and Smith (1972) and to the excellent discussion by Carlin and Louis (1996).

[31] We should note, however, that its possible to *over*-center in multilevel analysis. For instance, Bryk and Raudenbush (1992) suggest that all predictors should be centered, including dummy variables. In our view, this is counter-productive. The centering of dummies is not necessary either for the prevention of ill-conditioning or for the interpretation of results. In fact, we suspect that centered dummies hinder a clear interpretation of the results, rather than enhance it.