Political Analysis Advance Access published August 17, 2005

doi:10.1093/pan/mpi031

EDA for HLM: Visualization when Probabilistic Inference Fails

Jake Bowers and Katherine W. Drake
Department of Political Science, Center for Political Studies,
University of Michigan, Ann Arbor, MI 48109
e-mail: jwbowers@umich.edu (corresponding author)
e-mail: kwdrake@umich.edu

Nearly all hierarchical linear models presented to political science audiences are estimated using maximum likelihood under a repeated sampling interpretation of the results of hypothesis tests. Maximum likelihood estimators have excellent asymptotic properties but less than ideal small sample properties. Multilevel models common in political science have relatively large samples of units like individuals nested within relatively small samples of units like countries. Often these level-2 samples will be so small as to make inference about level-2 effects uninterpretable in the likelihood framework from which they were estimated. When analysts do not have enough data to make a compelling argument for repeated sampling based probabilistic inference, we show how visualization can be a useful way of allowing scientific progress to continue despite lack of fit between research design and asymptotic properties of maximum likelihood estimators.

> Somewhere along the line in the teaching of statistics in the social sciences, the importance of good judgment got lost amid the minutiae of null hypothesis testing. It is all right, indeed essential, to argue flexibly and in detail for a particular case when you use statistics. Data analysis should not be pointlessly formal. It should make an interesting claim; it should tell a story that an informed audience will care about, and it should do so by intelligent interpretation of appropriate evidence from empirical measurements or observations.
>
> —Abelson, 1995, p. 2

> With neither prior mathematical theory nor intensive prior investigation of the data, throwing half a dozen or more exogenous variables into a regression, probit, or novel maximum-likelihood estimator is pointless. No one knows how they are interrelated, and the high-dimensional parameter space will generate a shimmering pseudo-fit like a bright coat of paint on a boat's rotting hull.
>
> —Achen, 1999, p. 26

1 Introduction

Nearly all hierarchical linear models presented to political science audiences are estimated using maximum likelihood under a repeated sampling interpretation of the results

Authors' note: We owe many thanks to James Bowers, Andrew Gelman, Orit Kedar, Burt Monroe, Kevin Quinn, Phil Shively, Laura Stoker, Cara Wong, and the anonymous reviewers of the first version of our manuscript for their many helpful comments.

of hypothesis tests.[1] We all know that maximum likelihood estimators have excellent asymptotic properties but less than ideal small-sample properties. Multilevel models tend to have at least two different sample sizes. If an analyst has 10 countries with 1000 people inside each country, which sample offers the most guidance for assessing the appropriateness of standard maximum likelihood estimation of a multilevel model? If she is interested in the effects of country-level variables either alone or in interaction with individual-level variables, the relevant sample size is 10. Estimates from a multilevel model with 10 degrees of freedom for country-level variables may not be consistent and will not have known sampling distributions—and thus hypothesis tests in this case will be uninterpretable in the likelihood framework, even if those 10 countries were a random sample from the population of countries. While most analysts would realize that the asymptotic properties of a maximum likelihood estimator do not kick in at $N = 10$, many are understandably confused about a situation in which $n_j = 1000$ and $J = 10$—that is, where there is plenty of information for likelihood inference within countries but not enough for such inference across countries. This can be especially confusing if the dataset at hand has 10,000 rows (i.e., $N = n_j \times J =$ 10,000). Our anecdotal evidence suggests that this situation, in which the number of "containers" (like countries) is too small to support either powerful or credible hypothesis tests, is common in political science. For example, in a quick online survey of articles published in major political science journals from 2000 to 2005,[2] we found 14 articles that used multilevel models. Of those, half had level-2 sample sizes smaller than 28.

Does this mean that such a research design with only 20 countries is useless? We say no. In this article we argue that visualization of multilevel data can help analysts with small samples of level-2 units (like countries or states) learn about their data and present results that are not dependent on asymptotic or distributional assumptions. Although small samples mean that analysts have low power and/or uninterpretable hypothesis tests within the likelihood framework, small samples also have the benefit of being relatively easy to visualize. At small sample sizes audiences will prefer eleborate and focused description to simplification and summarization and so visualization appears particularly well suited to help analysts make the most of what data they have.

In other work we have emphasized the importance of designing multilevel studies such that the number of level-2 units is large (Stoker and Bowers, 2002a, 2002b), so we will not address research design directly here.[3] Snijders and Bosker (1999, p. 140) present a nice heuristic for deciding how to design a multilevel study when they say:

> A relevant general remark is that the sample size at the highest level is usually the most restrictive element in the design. For example, a two-level design with 10 groups, i.e. a macro-level sample size of 10, is at least as uncomfortable as a single-level design with a sample size of 10. Requirements on the sample size at the highest level, for a hierarchical linear model with q explanatory variables at

[1] When we say "hierarchical linear model" we mean a multilevel, mixed effects, random coefficients, or even random effects model. We will use the terms "multilevel model" and "hierarchical linear model" interchangeably throughout this article.

[2] *Comparative Politics, British Journal of Political Science, International Organization, Comparative Political Studies, Journal of Conflict Resolution, American Political Science Review, American Journal of Political Science, and Journal of Politics.*

[3] Other authors from a variety of disciplines have continued to emphasize the importance of large level-2 sample sizes. See, for example, the following recent papers and the citations therein: In a series of papers Hox and Mass present some simulations suggesting confidence intervals for level-2 coefficients are about 9% too small when $J = 30$ for a very simple model with one level-1 independent variable, one level-2 independent variable, and a cross-level interaction. They advise at least 50 groups in order to have correct coverage of confidence intervals on all pieces of the multilevel model (the fixed coefficients and the variance components) (Maas and Hox 2002, 2004; Hox and Maas 2002). Kreft (1996) reviews simulation studies and suggests a "30/30" rule (at least 30 level-2 units and at least 30 level-1 units within each of them).

> this level, are at least as stringent as requirements on the sample size in a single level design with *q* explanatory variables.

That is, if you would not trust a single-level maximum likelihood model with a given number of units, then you should not trust your multilevel model to estimate level-2 effects with that number of units either.

In addition, the common interpretation of likelihood inference requires that the level-2 units be a representative sample of some well-defined population. Political scientists do not tend to have such nice samples of common level-2 units like countries or other governmental entities. Much more often the level-2 units are the countries that have enough data to analyze; that is, they are a convenient sample, not a random one. Thus, even if one had enough countries to rely on the large-sample properties of likelihood inference, one might be hard pressed to interpret *p* values as generated from repeated sampling from some population.[4]

The point of this article, however, is not to harangue scholars about the dangers of small sample sizes and the unrepresentativeness of their samples or to stop the flow of research using multilevel models. Rather, it is to present a few techniques that can help scholars keep working, but to keep working with more confidence than is currently possible using multilevel models inappropriately. We call our recommendations "EDA" for exploratory data analysis because they rely on graphical presentations of the data rather than formal hypothesis testing. In this way, we shift the focus from inference about repeated samples of a well-defined population to compelling description of the patterns within a given dataset. This means that analysts can get about their work of learning about the world while presenting results that are not based on assumptions about asymptopia that they know from the outset are false in their dataset. We think that the techniques we present here allow analysts to tell compelling stories and to assess the implications of theories while remaining honest about what kinds of inference a given research design will bear. We do not suggest these techniques as a substitute for multilevel models when analysts do have large random samples of level-2 units from well-defined populations. But we have realized that, despite the common advice urging new research designs to collect as many level-2 units as possible that we cited above, there is little advice about what to do when the constraints of the research design are less than ideal. In other words, what should the analyst do if she desires to make inferences about level-2 effects (or cross-level interactions) but the number of level-2 units is small and even perhaps the level-2 units are not a representative sample from some well-defined population? To answer this question we present some ideas that we have gleaned from a number of places about how to continue to work even when large-sample likelihood-based inferences fail.

[4] We focus on what Rubin (1991) calls "repeated-sampling model-based inference" or what we call "repeated sampling" or "likelihood-based inference" in this article, because the majority of uses of multilevel models (or statistical models in general) in political science occur within this framework. One other common method for dealing with the problems posed by small, unrepresentative samples in multilevel models is to switch the meaning of "probability" from a repeated sampling, frequentist understanding to a Bayesian perspective. For example, although most of Raudenbush and Bryk (2002) deals with maximum likelihood approaches to hierarchical linear models, their chapter 13 presents a Bayesian example for a case with 19 level-2 units. The move to switch inferential frameworks can make the results of data analysis more meaningful and easier to interpret. However, as the number of level-2 units decreases, the influence of the choice of a particular prior (in this case, tending to be multivariate normal) and hyperpriors (independent uniform distributions) is larger. That is, less observed data makes the resulting posterior more sensitive to the priors—and thus makes sensitivity analysis for choice of prior that much more important. For a nice example of such a sensitivity analysis see Jackman (2004).

Our suggestions here are meant to help analysts deal with violations of the sample size assumptions necessary to support their inference. There are of course many other assumptions that undergird any given data summarization effort, including assumptions about probability distributions (i.e., is it reasonable to assume a normally distributed dependent variable and normally distributed coefficients), about the structural relationships (i.e., is a straight line a reasonable summary), and about omitted variables (i.e., is the association causal or not). In this article, we are mainly going to talk about the asymptotic and structural assumptions. Causal assumptions are best checked with sensitivity analysis (see, for example Rosenbaum, 2002, chap. 4) and distributional assumptions are best checked with diagnostics. The best way to assess the probability model is to first estimate a multilevel model and then follow the advice given in Pinheiro and Bates (2000, chap. 4), Gelman (2003, 2004), and Gelman et al. (2004, chap. 6).

We will first present the "standard" multilevel model, since it is this model that most analysts desire to use. Then we present a few plots that address the model—in effect, answering the questions posed by the standard multilevel model using visualization rather than probabilistic inference.[5] Finally, we suggest some general guidelines for when visualization may be most useful and when the simplification of modeling and probabilistic inference comes into its own.

2 The "Standard" Multilevel Model

When most people say "multilevel model" or "hierarchical linear model," they nearly always refer to a particular setup—a linear, additive, structural model, a normal probability model for the dependent variable and the error, and a multivariate normal probability model for at least some of the coefficients in $\boldsymbol{\beta}$. Here we explain this archetypal model so that we can know what kinds of graphical techniques might enable an analyst to approximate multilevel inference in cases in which there are too few level-2 units to support the usual maximum likelihood assumptions.[6]

2.1 *The Model*

Like any statistical model, a multilevel model involves two main components: (1) a structural model that specifies the functional form of the relationship among the variables, and (2) a stochastic model that uses probability distributions to encode information about how the values of the variables were produced in the world.[7] In the majority of uses of multilevel models, the structural model is linear and involves

[5] We highly recommend Cleveland (1993) for intelligent discussion for how probabilistic inference and visualization can interact and complement each other.

[6] This standard model is the one that is hard coded into the lme() command in R and Splus (Pinheiro and Bates 2000), into the HLM program (Raudenbush and Bryk 2002), and into most other common multilevel modeling routines. Other common programs for estimating this standard model include gllamm for Stata (http://www.gllamm.org) and proc mixed for SAS.

[7] Some people talk about the probability model as representing beliefs about the way that the observations depart from the structural model—that is, that the probability model refers to the distribution of $\boldsymbol{\varepsilon}$. Since $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ it is equally easy to talk about the probability model as referring to the way that the values for y were produced. In the multilevel model literature it is more common to talk about assumptions for the distribution of $\mathbf{y}$, and we will follow this practice here. In the end, any probability model about $\mathbf{y}$ can be re-expressed as a probability model about $\boldsymbol{\varepsilon}$, and vice versa.

interaction terms, such that for a simple model with two levels, one level-1 explanatory variable, X, and one level-2 explanatory variable, Z,

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij} \tag{1}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \nu_{0j} \tag{2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \nu_{1j}, \tag{3}$$

where $j = 1 \ldots J$ for the number of level-2 units and $i = 1 \ldots n_j$ for the number of level-1 units within a given level-2 unit. Here β_{0j} and β_{1j} are assumed to vary as if they were drawn from a multivariate normal probability distribution whose location (mean) is determined by the functions of Z_1 shown in Eqs. (2) and (3). The variance-covariance matrix of this multivariate normal is usually written with entries denoted by τ such that

$$\mathbf{\Sigma}_\beta = \begin{bmatrix} \tau_{11}^2 = \mathrm{Var}(\beta_{0j}) & \\ \tau_{21} = \mathrm{Cov}(\beta_{0j}, \beta_{1j}) & \tau_{22}^2 = \mathrm{Var}(\beta_{1j}) \end{bmatrix}.$$

Combining the previous three equations, we have:

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_{1j} + \nu_{0j} + (X_{ij})(\gamma_{10} + \gamma_{11}Z_{1j} + \nu_{1j}) + \varepsilon_{ij} \tag{4}$$

$$= \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{10}X_{ij} + \gamma_{11}Z_{1j}X_{ij} + (\nu_{0j} + \nu_{1j}X_{ij} + \varepsilon_{ij}). \tag{5}$$

Equation (5) requires that, for a given value of Z, a change of 1 unit in X has the same effect on Y whether the move is from 0 to 1 or from 99 to 100. This structural model also requires that the slope of the line representing the X, Y relationship varies at a constant rate across the range of Z.

One can also write this standard model combining the structural decisions with the stochastic ones using bold letters to denote matrices such that:

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \mathbf{\Sigma}_Y \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma}_Y) \tag{6}$$

$$\boldsymbol{\beta} \mid \mathbf{Z}, \boldsymbol{\gamma}, \mathbf{\Sigma}_\beta \sim N(\mathbf{Z}\boldsymbol{\gamma}, \mathbf{\Sigma}_\beta). \tag{7}$$

Nearly all analysts decide that the values of Y ought to be seen as arising from a process governed by the normal distribution. The extra variation in the intercept and slope parameters β_{0j} and β_{1j} is also nearly always understood as arising from a multivariate normal distribution.[8] Nearly always, the variance matrix of Y, $\mathbf{\Sigma}_Y$ is assumed to contain σ^2 for the variance of Y_{ij}, $\rho\sigma^2$ for the covariance of Ys within the same level-2 unit, and 0 otherwise.

The major benefit of this model is that it allows the analyst to specify directly the structure of her dataset in her statistical model and to estimate relationships taking this into account.[9] Equations (6) and (7) together allow us to write the analogue to

[8] In the Bayesian context, the parameters in γ are often given their own probability distributions (called "hyperpriors"), which themselves are governed by parameters fixed by the data analyst. See chapter 15 of Gelman et al., 2004 for more on the Bayesian perspective on these kinds of models.

[9] For much more discussion of the benefits, weaknesses, details, and implementation of multilevel models see Steenbergen and Jones (2002); Snijders and Bosker (1999); Kreft and Leeuw (1998); Longford (1993); Goldstein (1999); Pinheiro and Bates (2000); Raudenbush and Bryk (2002); Singer and Willett (2003); and McCulloch and Searle (2001).

Eq. (5)—showing how both X and Z combine to produce values of Y given the distributional assumptions such that

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_Y, \boldsymbol{\gamma}, \mathbf{Z}, \boldsymbol{\Sigma}_\beta \sim N(\mathbf{XZ}\gamma, \boldsymbol{\Sigma}_Y + X\boldsymbol{\Sigma}_\beta \mathbf{X}^T) \tag{8}$$

$$\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\beta, \boldsymbol{\Sigma}_\gamma, \boldsymbol{\Sigma}_Y \sim N\begin{pmatrix} (\mathbf{X}^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{y} + \boldsymbol{\Sigma}_\beta^{-1}\mathbf{Z}\gamma), \\ (\mathbf{X}^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1} \end{pmatrix}. \tag{9}$$

Equation (9) shows that the coefficients in $\boldsymbol{\beta}$ are an average of the within-unit regressions weighted by the variation within those regressions (See chapter 10 of Gill 2002 for more explanation of this model in the Bayesian context). Thus an important benefit of these models occurs when an analyst has *many* level-2 units with so little information in each unit that it is impossible to estimate coefficients with reasonable standard errors without pooling. This benefit is possible only, however, if the analyst is prepared to commit to pooling the data in this particular way, with these particular assumptions—which require many level-2 units.

In general, there are four ways to estimate the coefficients in Eqs. (5) or (8) and (9). The first way is to ignore the multilevel structure of the data and to estimate this model using OLS—using the same structural model and same probability model for Y but no probability model for the coefficients, which are assumed fixed. The approach ignores the fact that the error is not ε_{ij} but $(\nu_{0j} + \nu_{1j}X_{ij} + \varepsilon_{ij})$, which produces heteroskedasticity and serial correlation.

More important, it is not reasonable to assume that the units inside one level-2 unit are *exchangeable* with the units in another level-2 unit. Roughly, the fact that the level-1 units are not exchangeable means that it does not make sense to treat them all as if they arose from a common probability distribution. This means that one cannot write down a likelihood function as a simple product of identical distributions, and it also implies that the responses y_{ij} cannot be seen as arising independently of the level-2 unit within which they are nested. The very fact that an analyst wants to estimate coefficients for Z suggests that he does not believe the level-1 units are exchangeable. Exchangeability is a weaker property than independent identically distributed (iid), but it is a precondition for using probability distributions to pool information from disparate observations.[10] If the level-1 units are not plausibly exchangeable then they are not independent, the degrees of freedom available for hypothesis testing will be too large, and the hypothesis tests on the coefficients will be too liberal. One strategy that can help analysts out of these many problems is to include dummy variables for each level-2 unit in their equation, thereby only estimating coefficients conditional on each unit as required by the assumption of exchangeability; this dramatically increases the number of coefficients estimated. Fixed effects allow for the intercept to be different for each level-2 unit but for the slopes to be constant across units.[11]

The second approach is to collapse the data such that the mean of y within each level-2 unit $(\bar{y}_{\cdot j})$ is regressed on the mean of x within each level-2 unit $(\bar{x}_{\cdot j})$ and z. This approach implies that the analyst believes that the level-1 units are identical within level-2 units, and thus the mean provides as much information about y and x as the individual observations do. If this decision is not correct, then the analyst has needlessly

[10] See Gill (2002, chap. 10) and (Gelman et al. 2004, chap. 5) for accessible discussions of exchangeability in Bayesian data analysis. Regardless of the mode of inference, exchangeability is an important prerequisite for any use of a likelihood function.

[11] See Wooldridge (2002, chap. 10), Davidson and MacKinnon (1993, chap. 9.10), and Mundlak (1978) for more on fixed effects models.

thrown away a lot of information—and, more important, no longer has a model of an individual-level process. In the end, analysts must worry that the results of such a model reflect the process of aggregation more than, or instead of, an individual-level process (Achen and Shively 1995).

The third approach is to estimate the set of β_{0j} and β_{1j} separately for each level-2 unit. Each within-unit regression may satisfy the exchangeability (and iid) assumptions required for such models. This approach has the benefit of allowing the analyst to inspect the set of regression coefficients for heterogeneity, but it is not often a reasonable strategy for estimating the parameters in Eq. (5). The approach implies that the analyst believes that the level-2 units have nothing in common, and it tends to produce a lot of possibly noisy regression estimates that are hard to combine in a principled way.

The multilevel model approach attempts to combine the benefits of the second two approaches. By specifying a probability distribution for the coefficients (i.e., assuming that the coefficients themselves are exchangeable), the analyst can overcome the problem of the nonexchangeability of $\mathbf{y}$: $y_{ij} \mid \beta_j$ can be exchangeable and conditionally independent even if y_{ij} is not, as long as β_j is exchangeable. In addition, if it makes sense to assume that the coefficients are drawn from a common multivariate distribution, then sparse information inside one level-2 unit can be compensated for by information from the pooled sample overall. A final advantage of the multilevel model that is not shared by any of the previous ones mentioned above is that variance components such as τ_{11}^2 can be estimated. Researchers in education often care about these quantities because they can indicate the proportion of, say, child-level math achievement that can be accounted for by school-level policy manipulations versus child-level attributes.

From the standpoint of scientific inference, the within-unit regression (or the fixed effects regressions) produces results that only allow the analyst to generalize to those particular units, while the multilevel model—by thinking about the level-2 units as a sample from a population or as having values produced by a general stochastic process—allows for generalization to other level-2 units subject to that same process if the assumptions are correct. But in order to get the multilevel model off the ground, we must have enough level-2 units to make our probabilistic inferences feasible before we can be comfortable with generalization.

2.2 *Assumptions and Consequences*

So far, we have asserted that a small sample size causes problems for inferences based on multilevel models. But what exactly are the consequences? As we have discussed, the standard multilevel model relies for inference on the idea that the sample at hand is nearly infinite and that it was randomly selected from a well-defined population. That is, any problem faced by any other use of maximum likelihood for probabilistic inference is faced by the multilevel model. In the cases of multilevel model usage that are most common in political science, the second assumption, that the sample on hand is somehow representative of some population, is nearly always violated. However, violating this assumption alone is not enough to cause problems with the actual data analysis, just with the interpretation of the p values that form the basis of standard likelihood inference. That is, having a large unrepresentative sample of an ill-defined population merely makes it hard to believe the standard frequentist interpretation of the p values, not hard to believe that they arose from a sampling distribution whose asymptotic foundation is firm. The first assumption, though, is absolutely key to both estimation and interpretation of the model. Violation of the assumption that the sample at hand is nearly infinite negates the nice asymptotic properties of maximum likelihood estimates (MLEs) and places serious doubt on any hypothesis tests or confidence intervals.

If the sample size approaches infinity then the ML estimator's consistency properties tell us that our estimates will be "close" to the true value of the parameter in the population. Since the log-likelihood function is defined to be a sum of a series of independent random variables, the large sample allows the central limit theorem to give the analyst confidence that the sampling distribution of our estimate is approximately normal. We can use our knowledge of this fact to do hypothesis tests and construct confidence intervals. These properties are true regardless of the actual or the assumed distribution of our disturbances and are based merely on the central limit theorem (King 1989).

If the sample size is small, then, in addition to the iid assumptions (within a level-2 unit), the credibility of the distributional assumptions that are required for construction of the likelihood function become more important in justifying the standard likelihood hypothesis testing machinery. If the analyst is correct in assuming a normal distribution for $\mathbf{y}$ and a multivariate normal distribution for $\boldsymbol{\beta}$, then because these two assumptions produce one multivariate normal distribution that is a member of the exponential family, the ML estimators of the coefficients (γ) will have good small-sample properties, as they are functions of sufficient statistics. It is well known that when sampling from a normal distribution the ML estimator of σ^2 is biased downward in small samples even though it is consistent.[12]

However, if the sample size is small and the distributional assumptions are incorrect, then all bets are off. We depend on asymptotic normality to derive the distributions of hypothesis tests, such as the likelihood ratio or the Wald test. If we cannot assume asymptotic normality, we cannot be sure what distribution these tests follow and therefore what the *p*-values from them mean (King 1989). Without a large sample size, our ML estimates are potentially biased and our hypothesis tests lack power at best and are based on incorrect and unknown sampling distributions at worst. The large sample sizes of ML allow the analyst some wiggle room in specifying her likelihood function—a model of Poisson data based on a normal assumption might produce coefficients that are hard to interpret (i.e., we might think they represent a parameterization of μ in the normal pdf but they ought to represent the λ in the Poisson pdf), but a large enough sample will ensure that the formulas for our hypothesis tests and confidence intervals, which are based on asymptotically normal sampling distributions of estimates, are correct. Clearly applying the standard multilevel model to data with very few level-2 units in the hope of estimating and testing level-2 coefficients is incorrect.

Once a researcher decides that a multilevel model is an appropriate conceptual model for her data, she faces a number of decisions. First are decisions about exactly how the variables relate to each other. Do straight lines do a good job of summarizing the relationship between X, Z, and Y?

Second are decisions about the ways in which the values of the variables were produced. Is the normal distribution useful for thinking about $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta})$? Is it useful for thinking about $p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \gamma)$? Are these assumptions reasonable? Even if the slopes and intercepts "look like" they could have been generated by such a process, is it desirable to summarize them in this way? Is it useful to think of the values of $\boldsymbol{\beta}$ as all emerging

[12] See Greene (2002) for more discussion about the small-sample properties of single-level maximum likelihood estimators.

from one single distribution? (That is, if the values in **y** are only exchangeable conditional on **β**, are the values of **β** exchangeable? Or must yet more conditioning be specified?)

Finally, even if these above questions have been answered adequately such that the analyst believes Eq. (5) [or Eqs. (8) and (9)], she must ask whether she has enough data for credible and powerful hypothesis tests and consistent parameter estimates.

In what follows we present some very basic descriptive techniques to help justify decisions about exchangeability and linear, additive structure. We also attempt to use visualization to answer some of the questions posed by the multilevel model while using a small, nonrandom sample of level-2 units. As we noted earlier, assumptions about an appropriate probability model are best checked with diagnostics after running a multilevel model—and thus are more or less irrelevant to people who have too few units for confident estimation of this model. In the end, most of what is important to political scientists occurs in the structural model, and this is what receives the focus in the following pages.

Cleveland (1993, p. 12; emphasis in original) notes that visualization is a method for learning from data that is different from probabilistic inference.

> [Visualization] stresses a penetrating look at the structure of data. What is learned from the look is guided by knowledge of the subject under study. Sometimes visualization can fully replace the need for probabilistic inference. We visualize the data effectively and suddenly, there is what Joseph Berkson called *interocular traumatic impact*: a conclusion that hits us between the eyes.

When analysts do not have enough data to make a compelling argument for repeated sampling based probabilistic inference, visualization can be a useful way of allowing scientific progress to continue despite lack of fit between research design and asymptotic properties of maximum likelihood estimators.

3 An Application: Education and Political Participation

One of the most persistent and important findings in the political participation literature to date is that individuals who have more formal education are more likely to get involved in politics than those who have less. Discovering that education is a strong predictor of political participation has raised the question about what it is, exactly, that education does to facilitate political participation. The most recent answers to this question provide two mechanisms: education influences political participation via provision of "civic skills" and "civic status." These theoretical mechanisms have received empirical support from findings that individuals who have money, time, or organizational abilities—that is, people who have the civic skills to participate in politics—are those more likely to do so. And individuals who know a lot of other people, particularly politically active people—that is, people who have civic status—are also more likely to participate in politics than those who have less extensive and politically involved social networks.[13] Rosenstone and Hansen (1993, p. 76) summarize the distinction between the operation of skills and status succinctly:

> ...When political participation requires that knowledge and cognitive skills be brought to bear, people with more education are more likely to participate than people with less education. Participation, that is, requires resources that are appropriate to the task.

[13] The most recent and extensive articulation and defense of the "civic skills" and "civic status" points of view are provided by Verba et al. (1995) and Nie et al. (1996). Verba et al. (1995) coined the term "civic skills." We use the term "civic status" to describe the findings and argument of Nie et al. (1996). Huckfeldt (1979) presents an in-depth analysis of how status within social networks can influence political participation.

> On the other hand, education also indicates both the likelihood that people will be contacted by political leaders and the likelihood that they will respond. Educated people travel in social circles that make them targets of both direct and indirect mobilization. Politicians and interest groups try to activate people they know personally and professionally.

Our question is whether there are places where the relationship between education and political participation changes depending on the social context—and in particular on the educational context. In places where few people have college degrees (like West Virginia, where only 15% of the population aged 25 to 65 has a college degree), those who do have BAs ought to find that their educational status places them into more politically relevant social network positions—and thus they ought to be more advantaged by their education than people who are just one college educated person among many (like those living in Massachusetts, with 33% of the population college educated).[14] If education most strongly predicts political involvement in places with few highly educated people—like West Virginia—and only weakly does so in places with many highly educated people—like Massachusetts—then we might think that education is mainly acting to allocate politically relevant status to people. If an additional year of education provides the same boost to participation in all places, regardless of the educational inequality in the context, then we might think that education is mainly providing individuals with the skills necessary to overcome the costs of political participation. Of course, the first pattern of results might also suggest that education is providing skills that are politically relevant in only some places—that somehow political involvement in West Virginia is qualitatively different in terms of skills or status required than in Massachusetts. The second pattern might also indicate that politically relevant social status is structured in the same way in all of the places that we examine. However, either set of findings would contribute to the existing literature and, more important, suggest new avenues for research into how institutions and behavior interact to produce politics.

3.1 *Our Data and Approach*

In order to make our analysis slightly more realistic, we decided to worry about two potentially confounding variables. At the individual level, we decided to try to distinguish the effect of another year of education on political participation from the effect of gender [since we know that women tend to report less participation than men and also, among the older two-thirds of the population, tend to have less education than men (Burns et al. 2001)]. We also worried that the effect of state-level educational context on the individual-level relationship might be masking the effect of competitive races occurring in that state. That is, in addition to our primary trivariate causal relationship (where the effect of an individual's education on their political participation may vary by the educational context in which they live), we also allow for the presence of a potential confounding covariate at level 1 (the person's gender) and a covariate at level 2 [the competitiveness of the senatorial race(s) in that state in that year]. The reasoning for the level-2 covariate is that we want to distinguish between the effect of enduring features of the economic and social structure of the state (like the distribution of college education across the population) and the temporary features of a given election.

We represent educational context with data from the 2000 U.S. census on the percentage of the population in a state that has completed a college degree. The

[14] This is the argument of Nie et al. (1996), only they are more concerned about changes in the national educational context over time, rather than differences among places in educational context in one moment in time.

competitiveness of the 2000 election at the state level is represented with the difference in vote percentages received by the two major party candidates for the Senate. This variable runs from 0 (meaning that the two major party candidates had virtually the same percentage of the two-party vote) to 100 (meaning that there was either no election or no two-party contest such that the winner received all of the two-party vote).[15] Data on gender, individual years of education, and political participation are from the 2000 National Election Study (NES). We selected the 20 states from the NES that had the largest samples of survey respondents within them. This left us with 20 states containing between 31 and 135 respondents each with a total of 1247 respondents with valid responses about their education, political participation, and gender. Thus we have a small, nonrandom sample of states, each with what we are pretending is a random sample of individuals inside.[16] In this way we have a dataset that is similar to those common in political science but with perhaps fewer level-1 units per state. The size of the sample of states ($J = 20$) is too small for large-sample properties like consistency to provide credible bases for direct application of the multilevel model via MLE. If we had only two states, of course, we might not even need to use visualization, but 20 is enough to make it useful to summarize the within-state relationships somehow.

The analyses that follow are not meant to be an authoritative investigation of how social context changes the influence of education on political participation. However, they are an example of how one might do some of the intensive investigation of data that we think ought to precede all applications of multilevel models. We proceed here as if we believed the standard two-level probability model written in Eqs. (6) and (7). We represent the ideas about education and political participation as they usually are, with the following structural model:

$$\begin{aligned}
\text{Participation}_{ij} &= \beta_{0j} + \beta_{1j}\text{Education}_{ij} + \beta_{2j}\text{Sex}_{ij} + \varepsilon_{ij} && (10)\\
\beta_{0j} &= \gamma_{01} + \gamma_{02}\%\text{College Educated}_j + \gamma_{03}\text{Competitiveness}_j + \nu_{0j} && (11)\\
\beta_{1j} &= \gamma_{10} + \gamma_{11}\%\text{College Educated}_j + \nu_{1j} && (12)\\
\beta_{2j} &= \gamma_{20} + \nu_{2j} && (13)
\end{aligned}$$

Notice that Eq. (13) allows the effect of sex on political participation to vary across states but does not specify exactly what state-level variable governs this variation. The effect of education on participation, however, is assumed to vary as a function of the educational context of the state. Equations (10)–(13) can be combined into one equation analogous to Eq. (5).

$$\begin{aligned}
\text{Participation}_{ij} = {} & \gamma_{00} + \gamma_{01}\%\text{College Educated}_j + \gamma_{02}\text{Competitiveness}_j \\
& + \gamma_{10}\text{Education}_{ij} + \gamma_{20}\text{Sex}_{ij} \\
& + \gamma_{11}\%\text{College Educated}_j \cdot \text{Education}_{ij} \\
& + (\nu_{0j} + \nu_{1j}\text{Education}_{ij} + \nu_{2j} + \varepsilon_{ij}).
\end{aligned} \qquad (14)$$

[15] The largest two-party margin by which a contested Senate election was won in 2000 was 60 percentage points.

[16] We are aware of several additional complications and modifications that might be made to this model due to the clustering within states that arises from the general sampling design of the NES (Stoker and Bowers 2002a, 2002b) and specifics of the 2000 NES (Bowers and Ensley 2003). We set these concerns aside for this article since here the data play an illustrative role in the service of our discussion of methods.

The NES provided 11 questions about the nonvoting political involvement of respondents.[17] We summed the "yes" responses to these questions to create a variable containing the total number of acts that a respondent reported doing in the past year—only 50% of the sample reported more than one act. The NES respondents also reported their educational status, which produced a variable ranging from 0 years of formal education to 17 or more years—50% reported more than 14 years of education. Since only 39 people out of 1794 with valid education answers reported less than 8 years of education, we collapsed their answers to 8 years. In the analyses to follow we "centered" this variable such that $0=12$ years of education, $-4=8$ or fewer years of education, and $5=16$ or more years of education. The reason we did this is to ensure that the intercept (β_{0j}) has a meaningful interpretation (i.e., the average number of acts of political participation among men with a high school degree). As we shall see, certain states did not contain people with lower than a high school degree, so setting the zero point of the dependent variable to 12 years of education enables the intercepts of the within-state regressions to avoid displaying artifically high variance merely by virtue of an intercept outside the range of the data.

3.2 *Batches of Lines as Data*

The model as we wrote it in Eq. (14) specifies that the individual-level relationship between education and political participation ought to vary smoothly as a function of the percent of the state's population who have a college degree, holding constant the effects of gender and the competitiveness of the senate election. That is, the entity that we want to know about is the slope of the individual-level regression line. Since these slopes (and intercepts) are the objects of our substantive question, these are the objects that we would like to learn about—and thus they are the objects we ought to visualize. To this end we estimated a separate regression of participation on education and gender for each of the 20 states [in essence estimating the model shown in Eq. (10)].[18]

We collected the coefficients from the within-state regressions into a new dataset with one row per state and added the state-level variables. As noted above with the two probability models, $\boldsymbol{\beta}$ depends on some second-level variables, $\mathbf{Z}$, with some effect, $\boldsymbol{\gamma}$. The new data we created has our $\hat{\boldsymbol{\beta}}$, which we can think of as our dependent variable in the second level, as well as our $\mathbf{Z}$ (percent college educated in a given state and senate competitiveness), which we can think of as our independent variables at the second level.

If there is no appreciable variation in the slopes and intercepts across states, then it is possible that our model is incorrect in specifying that the individual-level relationships ought to be modeled as regressions, the slopes and intercepts of which are distributed according to a multivariate normal. That is, the first question we have is whether there is any appreciable variation in the within-state regression coefficients that is worth attempting to explain with a state-level variable. Pinheiro and Bates (2000) have developed a plot for just this purpose. For each state, they suggest plotting the point estimate with confidence intervals as line segments on either side. We present such

[17] The types of participation summed here are: Did R try to influence vote of others, Did R display button/sticker/sign, Did R go to meetings/rallies etc., Did R do any other campaign work, Did R contribute to candidate, Did R give money to party, Did R give to group for/against candidate, Worked on community issue in last year, Contacted public official to express in last year, Attended community meeting about issue in last year, Took part in protest or march in last year.

[18] This approach is the same as that suggested by Gelman for TSCS data, the "secret weapon" (see http://www.stat.columbia.edu/~cook/movabletype/archives/2005/03/the_secret_weap.html).

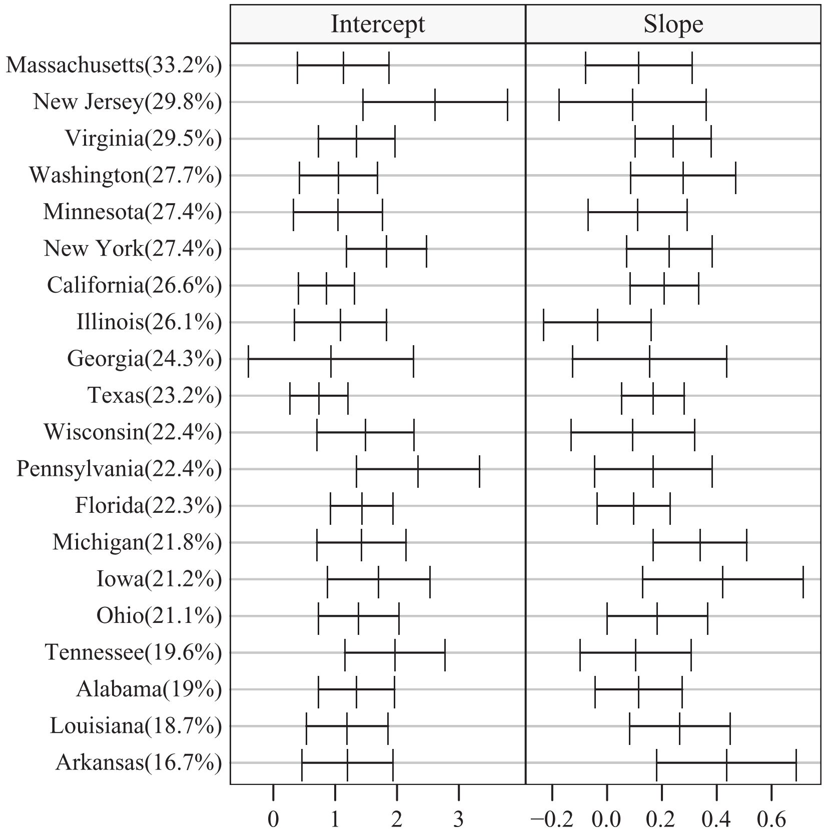


Fig. 1 The between-state variation of within-state regressions. The within-state regression estimates and 95% confidence intervals for the intercept and slope of education from Eq. (10) are plotted in order of state-level education (percent of the state population aged 25–64 who have at least a college degree).

a figure with 95% confidence intervals in Fig. 1. Even though we estimated a different effect for gender in each state such that these slopes on education represent the effect of individual education on individual participation controlling for gender, we do not display the coefficients for sex here, since we have assumed that it has the same effect across all states and since it is not the primary focus of our attention. The left panel displays the intercepts, which are to be interpreted as the average number of acts of nonvoting political participation reported by men in those states, since “male” is the excluded gender category. The right panel displays the slopes, which show the difference in average number of acts of political participation between people who have one year of education difference.

If all of these horizontal lines stacked up right on top of one another, then we might imagine that education has the same effect on participation in every state—and thus that the “education as social network status” mechanism may not be operating strongly (or that the relevant social networks are not as large or small as a state such that we have the wrong type of level-2 unit for this test). In fact, there is appreciable variability here (at least, we think that most reasonable scholars would think so). The differing widths of the confidence intervals merely suggest the differences in within-state sample size.

We plotted these lines in the order of state-level educational context, from the lowest at the bottom of the plot (Arkansas with 16.7% of the population aged 25–65 having at least a college degree) to the highest at the top of the plot (Massachusetts with 33.2% of the population having at least a college degree). If an individual's own education becomes more important in places where fewer people are well educated, then the lines in the "slope" panel of this plot should display a pattern where the bottom lines are in the right side of the plot (i.e., larger slopes) and the top lines are in the left side of the plot (i.e., smaller slopes). Such an overall pattern does not jump out of this plot. Notice that this plot did not take the competitiveness of the state Senate elections into account. However, it does suggest that while the individual-level relationship is not strictly constant across states, it does not vary extremely systematically by state-level education.

The plot in Fig. 1 does a good job of allowing us to compare the entities of theoretical interest (the within-state slopes and intercepts) to one another by simplifying them and presenting them side by side ordered by values of the state-level variable of interest. However, this plot does not allow us to assess the within-state fits. We already mentioned a concern with nonconstant residuals correlated with individual education; however, our model specifies OLS fits within states—and OLS is notoriously sensitive to being influenced by single points.[19] We would like to plot the within-state fits side by side, again in the order of our key second-level variable, to allow for assessment of how reasonable our decision to use OLS within states is. By ordering the plots using the state-level education we see how the within-state slopes vary as the state-level education changes—the previous plot was better for detecting this kind of pattern, and this next plot is better for assessing the within-state fits, but it still makes sense to keep the model in front of us as the baseline and motivation of our visualization. After all, our visualization is supposed to help us learn about the relationship between this model and our data.

Figure 2 shows the within state regressions from Eq. (10) with solid lines, outlier resistant within-state regressions with dashed lines, and the relevant bivariate scatterplots for each state.[20] The panels are plotted in increasing order of the state-level education context from bottom to top and left to right. The values of the state-level control variable, competitiveness of Senate race, are also noted within each panel. This particular display does not group states by competitiveness, so it is not well adapted to assess whether the state educational context interacts with competitiveness (we will assess this relationship in the next figure).

Each panel shows the scatterplot of the individual-level data for that state as gray dots. The black straight lines are the OLS fits—restricted to only plot within the range of the education of the individuals within that state. Thus the OLS fit for Arkansas runs from 8 to 17+ years of education (labeled here as -4 to 5 years of education from a high school degree), while the fit for Wisconsin is plotted from 11 years of education and up. The fact that all values of the within-state variable are not available for every state can be interpreted in two ways. First, it can be seen as evidence against our simple linear model—it might not be sensible to calculate some overall slope by averaging across these within-state fits if only certain states have

[19] See Langford and Lewis (1998) for a good discussion of how to detect influential outliers after running the full multilevel model. For a general discussion of the problem of influential points in OLS see Fox (1997, chap. 11).

[20] What we term an "outlier resistant regression" is also known as a "robust regression," where "robust" means "resistant to influential points." There is a large body of literature on robust estimation. The particular algorithm used here is called MM estimation (Yohai et al. 1991). We chose this particular algorithm because it has proven to be one of the most effective at resisting the effects of outliers; it works by combining two different robust regression algorithms (an S estimator followed by an M estimator using Tukey's biweight weighting function). For more details and citations on MM estimation and the implementation of the rlm function in R, see Venables and Ripley (2002, p. 161).

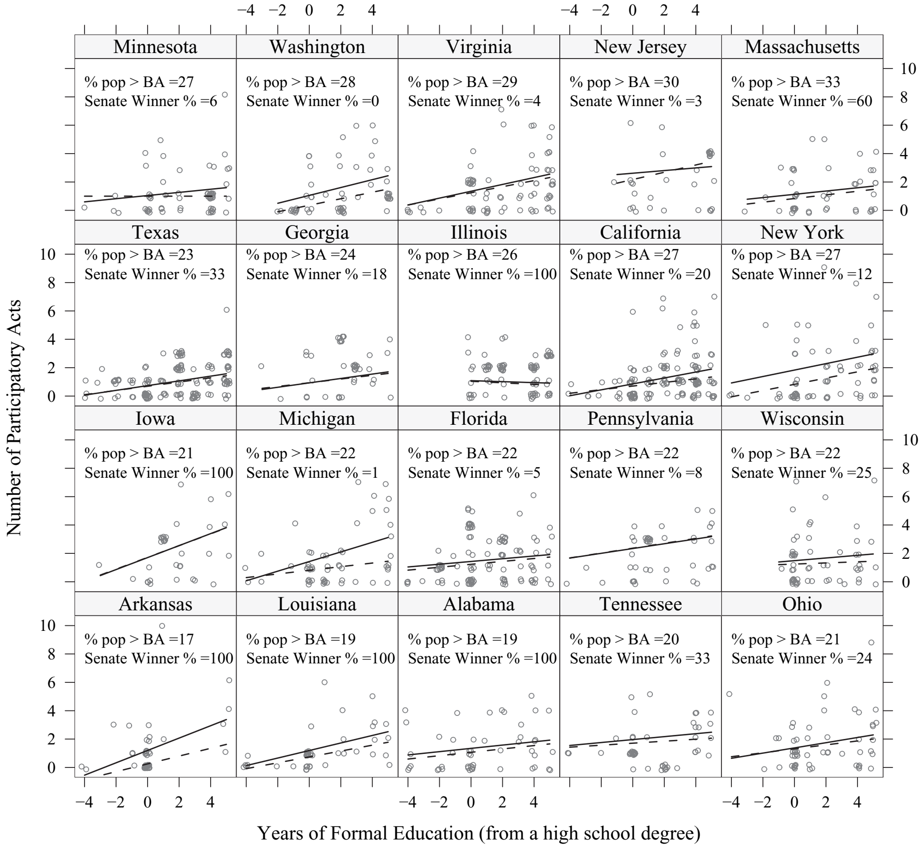


Fig. 2 Participation versus education within states. Points are slightly jittered to show density. Education less than 8 years is coded as −4, 12 years of education is coded as 0, education more than 16 years is coded as 5. The solid straight line is from OLS. The dashed straight line is from an outlier resistant linear model.

observed values on certain points of the scale. Another way to put this concern is in terms of exchangeability. Are all of these slopes representing the same stochastic process? What we are seeing here is that for Wisconsin and Illinois at least, the domain of this probability distribution appears different from the domain for the other states.

These plots do raise some questions about how the relationship between participation and education might be different among people with low education in this sample (especially in Wisconsin, New Jersey, and Illinois). Since inference about state-level education is driven by comparing the different slopes across the states, we have to be careful to realize that some of what we are comparing is actually missing. If we had many more states, this missing data problem could be dealt with rather elegantly by the multilevel model and the probability models that it assumes. However, if those probability models are incorrect, such that what is missing in Wisconsin, New Jersey, and Illinois ought not to be replaced with a function of what is going on in the other states, then the results of even a model based on a lot of data could be misleading.

This plot also shows that the classic ‘‘horn’’-shaped pattern of nonconstant variance is apparent in most of these panels. This suggests that we ought to be careful in interpreting standard errors for the within-state lines. But, since we are not interested in ‘‘telling the story of Arkansas" or any other particular state here, though, we are not worrying about this. In the context of many states and the standard multilevel model, however, this heteroskedasticity violates the standard within-unit probability model that we discussed in the section "The 'Standard' Multilevel Model."[21]

Each panel of this plot also contains a dashed line representing an outlier resistent fit. In five of the panels the resistent fit is easily distinguishable from the OLS fit. In the other panels it does not appear that any particular point is exerting undue influence on the fit. In each case in which the resistent line changes the fit, it moves the slope downward (thereby predicting less participation) and it tends to flatten the slope (thereby predicting a weaker individual-level relationship between participation and education). This kind of plot is also useful for detecting departures from linearity. In another version (not shown), we overlaid local linear (and quadratic) outlier resistent curves on the panels to compare with the outlier resistent regressions. We also overlaid lines connecting the means of the participation variable at each value of the education variable—this is the quantity that is being smoothed with the OLS line assumed in the model. We interpreted the results of those nonlinear local fits in comparison with the global linear fits (of OLS and outlier resistent regression) and did not detect any systematic patterns. The places where the lines curved the most tended to be in those ranges of individual education where the data were most sparse. Thus we interpreted the detectable nonlinearities as indicating places we were happy to smooth over rather than as illuminating critical features of the relationship that we were missing with the straight lines.

What is the substantive story from this plot? Does education appear to affect participation more in states where there are few college-educated inhabitants? The slopes of the OLS fits on the bottom two rows of this plot do not appear systematically steeper than those in the top two rows. Without controlling for the competitiveness of the particular election season, it does not look as though state-level educational context has much effect on whether highly educated individuals are more likely to get involved in politics.

When we look at Fig. 2, does a linear, positive relationship such as that represented with $\beta_{1j} = \gamma_{10} + \gamma_{11}\%$ college educated$_j + \nu_{1j}$ in Eq. (3) hit us between the eyes? To a seasoned user of regression, the evidence presented in this graph might not be very strong, especially given the few points in the upper right-hand corner of many panels of the plot that appear particularly well placed to exert undue influence on the slope of a globally linear fit (like a least squares line) within state. To deal with this concern the lines plotted in that figure are the result of outlier resistent fitting techniques. Thus the summaries presented by those lines are probably not artifacts of a few stray points. The fact that nonparametric local fits (not shown) and the global fits overlap quite a bit suggests that a straight line is a reasonable summary within the states—to the extent that the two global and local lines diverge it is relatively slight and in areas with relatively little data. Dealing with concerns about influential points and nonlinearity does not, however, answer the question about how to assess the results of data visualization. However, the process that we show using Figs. 1 and 2 embodies one general principle: that of comparison.

We find that visualization for multilevel models is most effective if it can allow the audience to rapidly and easily compare what is plotted to some baseline, or to other plots.[22] For example, we assessed linearity by comparing a straight line to a curved line. We checked the influence of outliers by comparing an outlier resistent line to a line fit with OLS.

[21] In a real substantive application this discovery would require us to revise the confidence intervals shown in Fig. 1, since they are based on the standard spherical errors assumption.

[22] The idea that we learn from comparison is not new. See, for example, Holland (1986) and Brady and Seawright (2004) for discussion of the fundamental role of comparison in establishing causal relationships.

By comparing the scatterplots in each panel to an imaginary plot where the points are evenly distributed around the line (easy to imagine and therefore not necessary to plot), we can see that the variance of the residuals will probably vary as a function of the education of individuals—that is, we will probably have within-unit heteroskedasticity problems that would need correcting before we felt comfortable with the standard errors arising from within-unit regressions. This principle that we learn by comparison is obvious, and it has been explained in relation to graphical displays of quantitiative data by Gelman et al. (2002) and the references cited therein. This also means that carefully describing the conceptual model that we believe to be a good representation of our substantive process is important—even if we know that we cannot easily estimate our model, or happily trust and interpret the results if we can easily estimate it. In this article the conceptual structural model is represented by Eq. (14)—it is our baseline, against which we will compare our displays.

Of course, reasonable people might still disagree about whether a particular line slopes upward enough to be distinguishable from flat. There is no easy answer to knowing what is "really there" in a given plot, and this article is not the place to grapple in depth with this problem.[23] If one has so little data that probabilistic inference is not easy, then the sizes of effects apparent in a given plot must be large—i.e., if we assume that the power to detect effects of a particular kind is more fine grained for the tools of probabilistic inference than for our eyes and substantive intuition, then the rule is essentially "we will know an effect when we see one"—and, to mix together yet another vague legal maxim, a "reasonable scholar" facing the same graph ought to agree that what is a feature of the graph is actually a feature. Reasonable people can disagree on the interpretation of features in the same way that they can disagree about the interpretation of coefficients, even if the numbers are somehow extreme enough to justify rejecting the common "flat slope" null hypothesis.

In what follows we will try to keep in mind (1) exactly what the baseline model is, such that we can know to look for features that we have said are meaningful when we specified the model (i.e., that we are not just looking for interesting patterns based on our substantive knowledge and intuition, but we are looking for evidence that speaks to our particular model), and (2) that when we detect a feature with our eyes, we will try to only report it as a feature rather than noise if we feel that any reasonable political scientist in our field would also detect this feature.[24]

Although we always begin with a plot like Fig. 2 in our own exploratory data analysis, it is most useful for assessing the reasonableness of the within-state fits and does not allow as direct a comparison of the units of analysis (the slopes) as did Fig. 1. However, that figure only allowed us to inspect the slopes capturing the relationship between individual-level education and individual-level participation controlling for gender by levels of state-level education. It did not allow any look at this relationship by the competitiveness of state-level elections. Our approach to addressing this question is to gather the slopes shown in the previous two figures into different panels of a single graph. By ignoring the individual

[23] We can, however, note that Buja and Cook (1999) recently proposed a method for inference from data visualizations that requires asking people who do not know much (or anything) about a particular piece of data analysis to pick the "most special looking" plot out of a group of N plots ($N - 1$ of which were generated using random draws from the probability model for a given null hypothesis of "no features" being apparent in a given graph). If the actual data plot is chosen from among the $N - 1$ other plots by a person, then one might say that this plot had a $1/N$ probability of being chosen merely through chance. Of course, documenting the results of such exercises would be difficult, but it is an interesting idea.

[24] There is much much more to say about data visualization than we can possibly cover in a short article devoted to examples rather than theory. Cleveland (1993) is a canonical reference, and we also recommend the work of Tufte (1983, 1990, 1997, 2003) as good places to start.

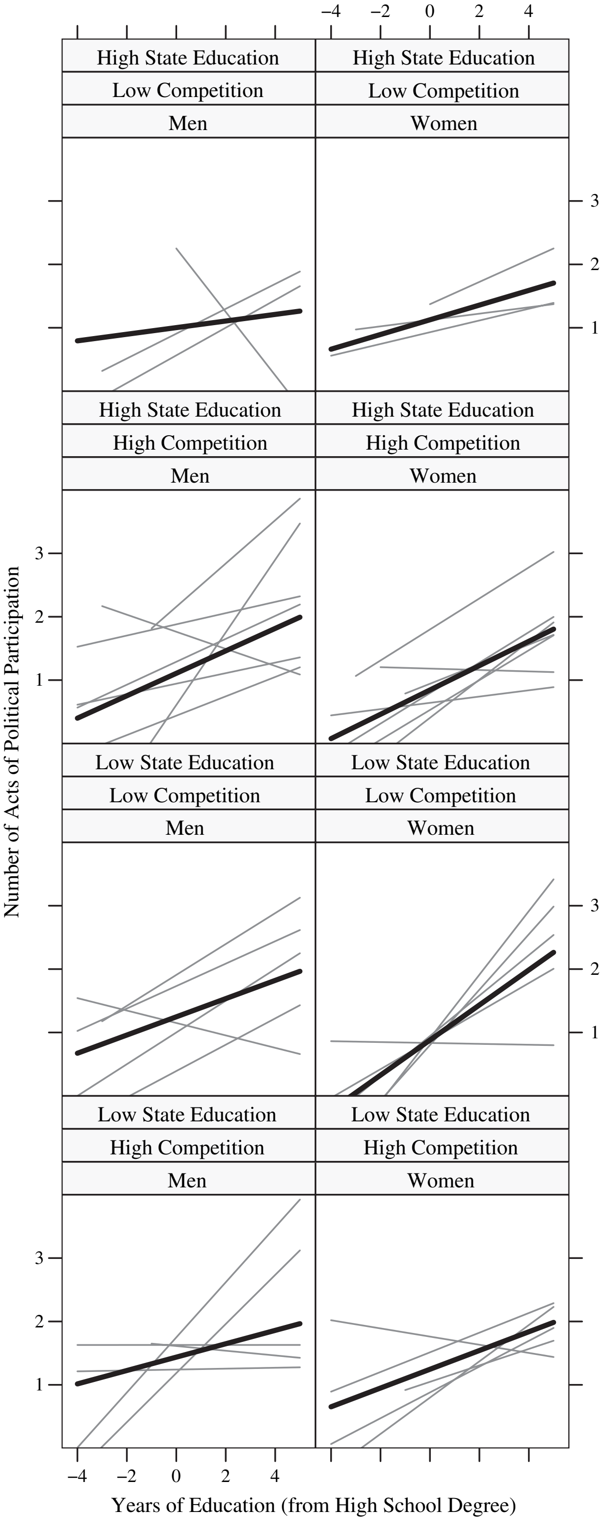
points and focusing on groups of slopes we can most directly assess the structural model that we specified in the beginning. In Fig. 3 each panel collects the regression lines for states with a given level of college-educated population split at the median—low (less than or equal to 22.5%) and high (greater than 22.5% college educated)—organized by level of Senate race competitiveness—also low (more than 25 percentage points separating the winner from the loser in the state Senate race) and high (less than or equal to 25 percentage points separating the winner and the loser in the state Senate race). The thick black lines in each panel have the mean intercept and mean slope of all of the lines in a given panel—weighted by the number of observations used in each within-state outlier resistent regression.[25] The grouping of these plots is also useful for a discussion of what it means to "control for" competitiveness and state-level education as well as individual gender. That is, some relationship estimated "controlling for" another variable is most easily understood as that relationship conditional on another variable remaining constant. As can be seen from the plot, the number of within-state slopes included in each category is not particularly high. If we added another level-2 covariate, we should begin to worry about the ability of the data to estimate such relationships—we would have to rely even more heavily on the linearity assumptions of OLS to "hold constant" the effects of these other variables.

Another way in which this plot addresses our original model is that the lines are no longer labeled. That is, the particular identity of a given line with a particular state is no longer apparent; instead what we see are collections of lines. This is what is implied by the use of a multivariate normal probability model for the slopes—that the particular identities of the units within which the slopes are identified are not important, but instead, an overarching relationship that somehow is a summary of the dynamics within all of the units is what is desired. In plots like this one we can see how this relationship varies by values of state-level variables.

If the educational context of a state changes the way that an individual's education influences her political participation, then we ought to see systematic differences in the slopes of the lines (either the thick ones or the thin ones) between the panels of this figure. This kind of pattern is not evident here—even the weighted mean lines appear to have nearly the same slope. This picture makes us rethink our initial ideas about state-level educational context as structuring or conditioning the way that individual education drives individual political participation. While our Fig. 1 suggests that there may be variation in the effect of individual education on participation across states, it seems from further investigation that our particular model does not capture the source of this variation.

One problem with Fig. 3 is that the number of lines per panel can be quite small. In part this is an unavoidable function of the small sample size of this study. However, in part this understates the amount of information in the two state-level variables—both are interval-level measures (or nearly interval level), not nominal. One way to show how a level-1 relationship can vary as a function of continuous level-2 variables is to transform the continuous variables into discrete variables of the traditional type, with

[25] We do not recommend this simple weighted average technique for inferential purposes. Lewis and Linzer (2005) have shown that the analogue to this technique, using simple within-place sample sizes in WLS of within-place regression coefficients on place-level variables, is neither consistent nor efficient and thus will produce misleading inferences. They recommend OLS with heteroskedasticity consistent standard errors or a version of FGLS for the second stage of such two-stage estimation of multilevel models. Since this article is about visualization as an alternative to probabilistic inference when the conditions of the research design cast doubt on the properties of single-stage multilevel estimators or on other estimators that require large level-2 samples (such as heteroskedasticity consistent standard errors), we merely use the weighted averages in this plot to reflect the fact that certain lines were calculated with more precision than others, not as components in probabilistic inference.


High State Education
Low Competition
Men
High State Education
Low Competition
Women
High State Education
High Competition
Men
High State Education
High Competition
Women
Low State Education
Low Competition
Men
Low State Education
Low Competition
Women
Low State Education
High Competition
Men
Low State Education
High Competition
Women
Number of Acts of Political Participation
Years of Education (from High School Degree)
−4
−2
0
2
4
1
2
3

mutually exclusive categories. This is what we did in Fig. 3. However, one can also transform the continuous variable into, say, two or three pieces, which have overlapping categories. This increases the number of within-state lines in any given panel and provides a better sense for how relationships may change smoothly (or not) across the domain of the continuous variables—rather than perhaps being an artifact of the particular cut point chosen in the mutually exclusive category transformation used earlier.

Figure 4 shows such a plot. Each panel contains the within-state outlier resistent regression lines in gray and a weighted average line overlaid. The bottom nine panels show the relationship for men, and the top nine show the relationship for women. Each of the two continous state-level variables has been broken into three overlapping pieces (called "shingles" because they overlap): percent of the state population college educated has three pieces (16.7% to 23.2%; 21.8% to 27.4%; 23.2% to 33.2%); competitiveness of the state Senate race also has three pieces (0 to 12.5 percentage points; 7.5 to 33.5 percentage points; 32.5 to 100 percentage points). These shingles were chosen to have nearly equal numbers of survey respondents in each group and to overlap somewhat, but not too much. The bottom row of the plot shows how the individual relationship among men changes as the state-level education changes—holding competitiveness constant at the most competitive level (races where the winner won by less than 12.5 percentage points of the two-party vote). Reading from the bottom of the first column up three rows, we can see how the individual-level relationship relates to changes in the competitiveness of the Senate race, holding state-level education constant at 16.7% to 23.2% of the population having a college degree.

Although the panels are still sparse, because we have allowed the repetition of states across adjacent panels, we have more information that we can use to examine the multilevel relationship. However, this plot does not make the relationship between educational context and the individual-level relationship between education and political participation more clear: the lines do not become more or less flat in any systematic way across the values of these variables. Some of the individual-level relationships (under certain combinations of values of gender, competitiveness, and percent college educated) appear slightly more or less strong than others. Overall, comparing overlapping groups of states is useful insofar as it does not arbitrarily discretize what are continuous variables, and insofar as it allows an analyst to look for a changing relationship across neighboring panels.

In this particular example, these plots have not supported the model that we specified in the beginning. Does this mean that the method of visualization is useless? Of course not. The results that we have shown are analogous to having shown a MLE multilevel model with coefficients where the relevant hypothesis tests do not allow rejection of the null. Given the strong priors about how education ought to matter differently in different

⟵

Fig. 3 Within-state outlier-resistent regressions grouped by levels of state-level educational context. The thin gray lines show the within-state, outlier-resistent regressions (MM estimates) grouped by (1) gender of the respondent, (2) competitiveness of the state Senate race (less than or equal to 25 percentage points difference = high competition, greater than 25 percentage points difference = low competition), and (3) percent of the state population college educated (median split at 22.5% with a college degree or more). The thick black lines were generated from averaging the intercepts and the slopes of the gray lines, weighting the average by the sample sizes within state. Unweighted averages were not detectably different except for men in high education states where the Senate races were not very competitive.

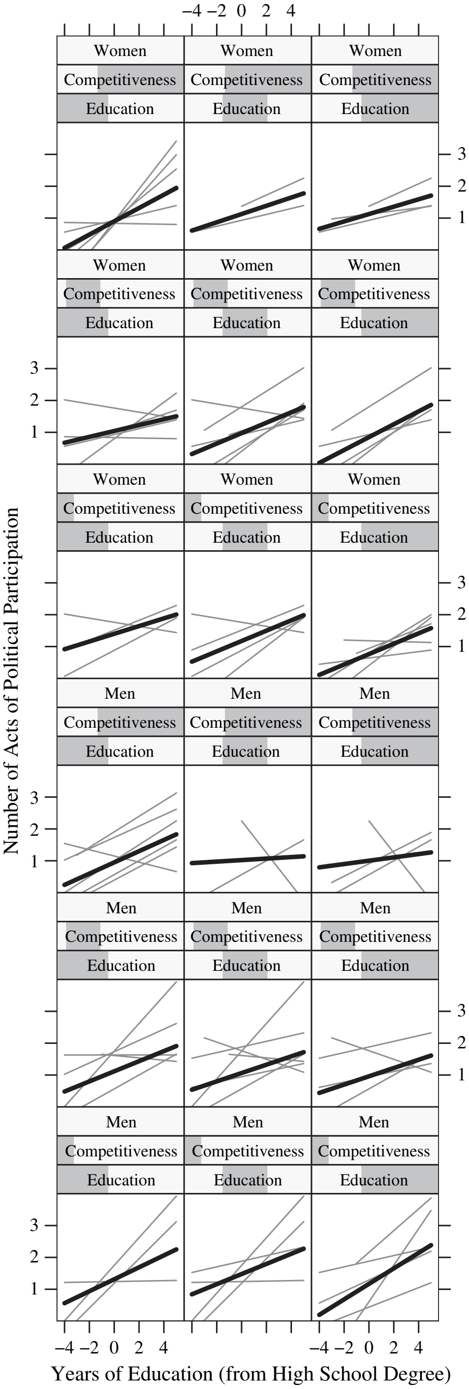


Fig. 4 Conditioning with two continuous variables using shingles. Within-state individual-level outlier resistent regressions of participation on education are shown in gray in each plot. The black lines are the averages of the gray lines, weighted by the sample sizes of each state. The panels are ordered from bottom to top and left to right in order of levels of state-level education (low to high going left to right—the part of the education scale portrayed in each plot is shown by the shaded bars

educational contexts, this is an example where visualization has provided an interesting substantive result.

We have learned that the relationship between political participation and years of education among individuals in 2000, among these states, is not strongly related to the educational context in which people live. This result argues against the idea that education operates mainly to provide politically relevant social status to people. If an additional year of education appears to provide more or less the same amount of participatory advantage across places with different educational contexts, then perhaps (1) education in the contemporary United States mainly matters to enhance political activity via the provision of skills or some other attribute of individuals that can affect political activity more or less the same across the nation, or (2) the kind of status allocation provided by education is not well characterized by state-level variables but ought to be measured instead by, say, relative standing within one's birth cohort [which is what Nie et al. (1996) do] or some combination of birth cohort and geography [which is what Tenn (2005) does]. Or perhaps politically relevant social network position is national now. It does not look as though it is operating at the state level, at least among these states, even controlling for the short-term effects of Senate campaigns. Notice that we are making these claims with some humility due to our small sample size, but we are not underselling our results. If this were the only dataset available on this topic, then these results without a hypothesis test would be the best we could do without switching modes of inference.

If we were to implement a multilevel research design with many places (say, towns) sampled at random from the nation, with representative samples of individuals within them, the results from this study might suggest that the standard multilevel model might not be appropriate, even if we had enough level-2 units to engage confidently in likelihood-based inference. Nonconstant variance and influential points within places would probably be a concern. And one might think that assuming simple cross-level interaction effects as we did here would be either misleading or incorrect. We might prefer to investigate the relationships between educational context and respondents' education and participation more intensively—perhaps allowing a much less smooth relationship between them than that assumed by the simple $\gamma_{11}Z_{1j}X_{ij}$ in Eq. (5).

Of course the example that we presented here has been quite brief. However, we hope it has served to illustrate a data analysis that relies on visualization rather than on repeated-sampling based probabilistic inference. But merely illustrating visualization in this way does not provide much of an easy to articulate rationale. The main point of this article is that when analysts do not have enough data to make a compelling argument for repeated-sampling based probabilistic inference, then visualization can be quite useful; it does not allow scientific progress to halt because of a lack of fit between the research design and the asymptotic properties of maximum likelihood estimators.

4 Why Graph?

We have suggested that analysts who do not trust the assumptions required for probabilistic inference do not have to throw away their data, but can fall back on

←

labeled "Education"), then levels of competition in the state Senate race (going from very competitive to not very competitive from bottom to top—the part of the competitiveness scale portrayed in each panel is shown by the shaded bars labeled "Competitiveness"). The bottom three rows show the results for men and the top three rows show the results for women.

visualization as a tool that can allow research to progress—even if the strength and precision of arguments made on the basis of visualization are less than those based on large, representative datasets. However, when a result does emerge from visualization, it does hit the audience between the eyes, and thus may be as compelling as many asterisks beside a coefficient in a table.

Another important reason to visualize, even in the presence of plenty of data, is to check assumptions. Cleveland (1993, p. 14) summarizes the argument for this: "Without a careful checking of assumptions, validity is replaced by large leaps of faith; one can only hope that the probabilistic assertions of confidence intervals and hypothesis tests are valid. Visualization reduces leaps of faith by providing a framework for studying assumptions." We do not elaborate on the assumption checking reason for visualization here, mainly because there is good literature on this topic, even if it is not common to see political scientists use this advice in practice.[26]

Another reason that visualization is especially useful for small samples is that the value of an additional piece of information to readers increases in the range of sample sizes that we are talking about here. That is, most people would agree that when it comes to scientific communication, more information is usually better. This intuition, for example, is behind much of Edward Tufte's recent criticism of Powerpoint (Tufte 2003). The problem with Powerpoint and other screen-based presentation software, from Tufte's point of view, is that, by design, they decrease the amount of relevant information that a speaker can (or ought to) convey to an audience. Each screen in Powerpoint tends to contain a small number of bullet points that delimit speaking points. In that polemic Tufte advocates handouts on the basis that the printed page can carry a much higher density of information than the screen of a single slide, and notes that more information in fact can enhance an argument rather than hurt it: "Often, the more intense the detail, the *greater* the clarity and understanding—because meaning and reasoning are *contextual.* Less is a bore" (Tufte 2003, p. 10; emphasis in original).

One theme that Tufte has stressed in his writing about both presentation and display of quantitative data is that good analysis and presentation are akin to good teaching (Tufte 2003, p. 11); it is the opposite of the sleight of hand that characterizes magic (Tufte 1997, pp. 68–71). In fact, plots of within-unit fits are a common feature in the textbooks on multilevel modeling that we listed in note 9; they are quite often useful in motivating what it is that such models actually do (i.e., produce a coefficient estimate that is in some way an average of the within-unit coefficients weighted by the amount of information contributed by each unit). However, such teaching devices are not commonly found in the working data analyst's tool kit, which is strange given that at least two of those textbooks explicitly include chapters on visualizing and describing multilevel data (Pinheiro and Bates 2000; Singer and Willett 2003).

Of course, there is no direct way to present a dataset with 200 countries and 1000 individuals inside each one. In that case, even some of the plots that we have shown here would be awash in a jumble of lines, and we would beg the analyst to model rather than describe. This distinction between the moments when we as readers would feel more confident in a presentation of statistical data if we only saw more detail, and those moments when we need summarization and guidance, suggests a simple relationship between the amount of information available in a given dataset and the marginal value of presenting pieces of that information. As, say, the number of countries in a multilevel

[26] See, for example, Langford and Lewis (1998); Gelman (2003, 2004).

dataset increases from 1 to 30 or even 50, we as readers would be grateful for more detail—in this range we desire elaboration and description of the cases. However, plots with more than 50 panels, or more than 50 lines, can be overwhelming. After this point, then, the smoothing and simplification of a model is more desirable. By presenting only coefficients and asterisks for relatively small samples—rather than informative tables or plots (or table graphics like that shown in Tufte 2003, p. 16)—analysts would be withholding useful information from their audience. When the sample size is small, and especially when these coefficients are the end result of myriad modeling decisions, audiences ought to be more skeptical, and thus more information and detail is needed. At some point, of course, the sheer number of within-place lines, points, and table entries can be overwhelming; when that happens, modeling ought to take the place of description.

The convenience in thinking about information presentation as having value almost like a commodity is that it seems as though the desires of the consumers of scientific information tend to shift from elaborate description to simplified models, roughly parallel with the rates of convergence of large-sample properties of maximum likelihood estimators. That is, one sign that more information ought to be presented is concern about the large-sample properties of estimators. Thus perhaps the requirements of scientific communication may map onto some of the requirements of the most common mode of probabilistic inference in political science.

In general, we think that taking a close look at the micro-level relationships within macro-level units enables analysts to demystify what can often seem like the black box of multilevel models. In addition, inspecting such data displays calls on the substantive knowledge and judgment of the scholar. And good judgment about modeling decisions is exactly what exploratory data analysis is about. If in pursuit of such good judgment unexpected patterns emerge, so much the better.

Of course, the plots and ideas presented here are not meant to be a particular set of techniques to be applied everywhere. They are meant to stimulate scholars to develop their own data displays, and we have cribbed many of them from the scholars we have cited. The basic idea is that multilevel models are partly about understanding patterns in $\mathbf{y}$ as a function of $\mathbf{X}$ but are also about understanding patterns in $\boldsymbol{\beta}$. Scholars are used to exploring the relationships within their datasets using crosstabulations and bivariate scatterplots but are perhaps not as used to exploring the relationships between their coefficients. In this article we have tried to suggest and present a few different ways that data analysts can make their varying coefficients easier to handle.

In an article about tools, it is worth giving credit to the tool makers. The plots shown in this article were all created with the R language (R Development Core Team, 2005). Within R they all relied on the Trellis graphics system that was specifically designed for visualization of conditional relationships (Becker et al. 1996; Cleveland 1993). The implementation of Trellis graphics in R used here is from the Lattice package (Sarkar 2005). All of the code used to produce this article is available embedded within the source of the document itself in Sweave format at http://www.umich.edu/~jwbowers/papers.html.[27]

> A basic problem about any body of data is to make it more easily and effectively handleable by minds—our minds, her mind, his mind. To this general end:
>
> - anything that makes a simpler description possible makes the description more easily handleable.
> - anything that looks below the previously described surface makes the description more effective. (Tukey 1977, p. v)

[27] For details on Sweave see Leisch (2002, 2005).

References

Abelson, Robert. 1995. *Statistics as Principled Argument.* New York: Lawrence Erlbaum.

Achen, Christopher H. 1999. "Warren miller and the future of political data analysis." *Political Analysis* 8: 142–146.

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

Becker, R. A., W. S. Cleveland, and M. J. Shyu. 1996. "The visual design and control of Trellis Display." *Journal of Computational and Statistical Graphics* 5:123–155. (Available from www.stat.purdue.edu/wsc/papers/trellis.design.control.ps.)

Bowers, Jake, and Michael Ensley. 2003. "Issues in Analyzing Data from the Dual-Mode 2000 American National Election Study." Technical Report. Ann Arbor, MI: National Election Studies.

Brady, Henry, and Jason Seawright. 2004. "Framing social inquiry: From Models of Causation to Statistically Based Causal Inference." Working paper.

Buja, Andreas, and Dianne Cook. 1999. "Inference for Data Visualization." Presented at the Joint Statistics Meetings, August 1999. Baltimore, MD. (Available from www-stat.wharton.upenn.edu/buja/PAPERS/jsm99.ps.gz.)

Burns, Nancy, Kay L. Schlozman, and Sidney Verba. 2001. *The Private Roots of Public Action: Gender, Equality, and Political Participation.* Cambridge, MA: Harvard University Press.

Cleveland, William S. 1993. *Visualizing Data.* Summit, NJ: Hobart.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods.* Thousand Oaks, CA: Sage.

Gelman, Andrew. 2003. "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing." *International Statistical Review* 71:369–382.

Gelman, Andrew. 2004. "Exploratory Data Analysis for Complex Models (with Discussion by Andreas Buja and Rejoinder)." *Journal of Computational and Graphical Statistics* 13:755–787.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. 2002. "Let's Practice What We Preach: Turning Tables into Graphs." *Statistical Computing and Graphics* 56:121–130.

Gill, Jeff. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach.* Boca Raton, FL: Chapman and Hall/CRC.

Goldstein, H. 1999. *Multilevel Statistical Models.* London: Edward Arnold.

Greene, William H. 2002. *Econometric Analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association,* 81:945–960.

Hox, J. J., and C. J. M. Maas. 2002. Sample Sizes for Multilevel Modeling. In *Social Science Methodology in the New Millennium. Proceedings of the Fifth International Conference on Logic and Methodology*, eds. J. Blasius, J. Hox, E. de Leeuw, and P. Schmidt. Opladen, Germany: Leske + Budrich Verlag.

Huckfeldt, R. R. 1979. "Political Participation And The Neighborhood Social Context." *American Journal of Political Science* 23:579–592.

Jackman, Simon. 2004. "Bayesian Analysis for Political Research." *Annual Review of Political Science* 7: 483–505.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* New York: Cambridge University Press.

Kreft, Ita. 1996. "Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies." Unpublished manuscript.

Kreft, I., and J. D. Leeuw. 1998. *Introducing Multilevel Modeling.* London: Sage.

Langford, Ian H., and Toby Lewis. 1998. "Outliers in Multilevel Data." *Journal of the Royal Statistical Society A* 161:121–160.

Leisch, Friedrich. 2002. Dynamic Generation of Statistical Reports Using Literate Data Analysis. In *Compstat 2002—Proceedings in Computational Statistics*, eds. W. Haerdle and B. Roenz. Heidelberg, Germany: Physika Verlag, pp. 575–580.

Leisch, Friedrich. 2005. "Sweave User Manual." (Available from www.ci.tuwien.ac.at/leisch/Sweave.)

Lewis, Jeffrey B., and Drew A. Linzer. 2005. "Estimating Regression Models in Which the Dependent Variable Is Based on Estimates." *Political Analysis.* doi:10.1093/pan/mpi026.

Longford, N. T. 1993. *Random Coefficient Models.* Oxford: Clarendon.

Maas, Cora J. M., and Joop J. Hox. 2002. "Robustness of Multilevel Parameter Estimates against Small Sample Sizes." In *Social Science Methodology in the New Millennium*, eds. J. Blasius, J. Hox, E. de Leeuw, and P. Schmidt. Opladen, Germany: Leske + Budrich. (Available from www.fss.uu.nl/ms/jh/papers/p090101.pdf.)

Maas, Cora J. M., and Joop J. Hox. 2004. "Robustness Issues in Multilevel Regression Analysis." *Statistica Neerlandica* 58:127–137.

McCulloch, Charles E., and Shayle R. Searle. 2001. *Generalized, Linear, and Mixed Models.* New York: John Wiley and Sons.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46:69–85.

Nie, Norman, Jane Junn, and Kenneth S. Barry. 1996. *Education and Democratic Citizenship in America.* Chicago: University of Chicago Press.

Pinheiro, José C., and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS.* New York: Springer-Verlag.

R Development Core Team. 2005. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. (Available from www.R-project.org.)

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Thousand Oaks, CA: Sage.

Rosenbaum, Paul R. 2002. *Observational Studies.* New York: Springer.

Rosenstone, Steven, and John M. Hansen. 1993. *Mobilization, Participation and Democracy in America.* New York: MacMillan.

Rubin, Donald B. 1991. "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism." *Biometrics* 47:1213–1234.

Sarkar, Deepayan. 2005. Lattice: Lattice Graphics. R Foundation for Statistical Computing [producer and distributor]. (Available from http://cran.r-project.org/src/contrib/Descriptions.lattice.html.)

Singer, Judith D., and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Snijders, T., and R. Bosker. 1999. *Multilevel Modeling: An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage.

Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46:218–237.

Stoker, Laura, and Jake Bowers. 2002a. "Designing Multi-level Studies: Sampling Voters and Electoral Contexts." *Electoral Studies,* 21:235–267.

Stoker, L., and J. Bowers. 2002b. "Erratum to 'Designing Multi-level Studies: Sampling Voters and Electoral Contexts'." *Electoral Studies* 21:535–536.

Tenn, Stephen. 2005. "An Alternative Measure of Relative Education to Explain Voter Turnout." *Journal of Politics* 67:271–282.

Tufte, Edward. 1983. *The Visual Display of Quantative Information.* Cheshire, CT: Graphics.

Tufte, Edward. 1990. *Envisioning Information.* Cheshire, CT: Graphics.

Tufte, Edward. 2003. *The Cognitive Style of Powerpoint.* Cheshire, CT: Graphics.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative.* Cheshire, CT: Graphics.

Tukey, John W. 1977. *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S-PLUS*, 4th ed. New York: Springer.

Verba, Sidney, Kay L. Schlozman, and Henry Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics.* Cambridge: Harvard University Press.

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

Yohai, V., W. A. Stahel, and R. H. Zamar. 1991. A Procedure for Robust Estimation and Inference in Linear Regression. In *Directions in Robust Statistics and Diagnostics, Part II*, eds. W. A. Stahel and S. W. Weisberg. New York: Springer-Verlag.