Designing multi-level studies: sampling voters and electoral contexts

Laura Stoker * , Jake Bowers

Department of Political Science, 210 Barrows Hall, University of California, Berkeley, California 94720-1950, USA

Abstract

Every two years in the U.S., 435 congressional elections take place that scholars study using data from the National Election Studies (NES) survey of the American electorate. With a focus on sampling, this article explores two issues: (1) How best to design a national election study if the aim is to understand voting behavior within and across subnational contexts; and (2) How, by comparison, the existing NES surveys have been designed. Although our arguments specifically address how one should sample individuals and congressional districts in the U.S., our conclusions apply to any situation where one is sampling micro-level units nested within diverse and influential macro-level contexts.

1 Introduction

Data from national election studies are often used to study voting behavior in subnational electoral districts. In the United States, the National Election Studies (NES) surveys the American electorate every two years, collecting data relevant to explaining election outcomes at the national (presidential) and subnational (congressional) level. Although national survey data are valuable for studying each type of election, the fact that congressional elections are inherently subnational events has implications for survey design. In any given election year, 435 different contests take place across the country, in districts with varying characteristics, with varying pairs of candidates running varying election campaigns. Hence, the questions we address in this paper arise: How should one design a *national* election study in order to best understand these

* Corresponding author: Tel. +1 (510)643-6872
Email address: stoker@socrates.berkeley.edu (Laura Stoker).

diverse *subnational* contests? And, how, by comparison, have the American National Election Studies been designed?

These are large questions, and, if left unfocused, larger than we can confront in this paper. We focus, therefore, on one crucial aspect of the design of a national survey: the sample design. Although sampling considerations are important to the design of any survey, they are particularly important in the kind of case we have at hand—where individuals from all across the nation are surveyed concerning the particular elections taking place within the districts in which they reside. In order to analyze these elections fruitfully, scholars must link micro-level survey data to macro-level data on attributes of the districts, the candidates, and the campaigns. As we will demonstrate, the effectiveness of the resulting analysis depends, to an important extent, on how the survey sample is designed.

In what follows we develop our arguments with respect to the problem of sampling citizens within and across congressional districts in the United States. The sampling issues we confront, however, are much more general than this. They arise any time that one is sampling micro-level units nested within diverse and influential macro-level contexts. They arise, for example, for researchers designing survey research on voting in U.S. presidential elections, in that presidential campaigns are waged very differently across the 50 U.S. states (Shaw, 1999a,b). If that state-by-state variation influences how voters make their choices, then researchers must confront how to sample voters within and across states. In this case, as with the case of congressional electoral research that we address in detail, below, sampling designs constrain what one will be able to learn. We first elaborate this point through example, by considering two contrasting survey designs.

2 Two Contrasting Designs

Imagine a design where 3000 eligible voters are selected at random within a single congressional district (CD), which, itself, is selected at random from within the 435 congressional districts in the U.S. With these data, we could study how various individual-level explanatory variables affect a dependent variable like vote choice in this election, although we could not draw conclusions from such findings about congressional elections in general. More importantly, we could not study how district, candidate, or campaign characteristics influence how citizens vote, since the study design generates no variation on such variables. For the same reason, we could not study how CD-level characteristics interact with individual-level characteristics in influencing the vote. As a consequence, our model of the vote will be incomplete, and hence inadequate for explaining the election outcome even within the congressional

district we studied. In this design, any effect of CD-level variables is operating in a wholly unobserved fashion. Even with data on 3000 eligible voters, we could not fully explain why people voted, why they voted the way they did, or why candidate A ended up beating candidate B. [1]

By contrast, imagine a design in which 100 congressional districts are first sampled at random, and then within each CD 30 eligible voters are randomly selected. [2] Like the first design, we could use the sample of 3000 respondents to evaluate the effects of individual-level characteristics like partisan identification. Since this design builds in variation at the CD-level, we could also study how district, campaign, and candidate characteristics influence the vote (assuming we gather the requisite data at the CD-level). For the same reason, we could study the interaction of CD- and individual-level explanatory variables, asking, for example, whether the effect of party identification depends upon the nature of the campaign. Since respondents are sampled at random within districts and the number per district is relatively large ($n = 30$), we could also aggregate responses within districts to generate contextual independent variables (e.g., "climate of opinion" variables), or could link the aggregated survey findings to data on members of Congress, as did Miller and Stokes (1963) in their study of representation. We could even draw district-level conclusions about the behavior of voters or about election outcomes, although the within-district n of 30 is limiting in this respect. However, since both districts and respondents are sampled at random, we could pursue the same kind of exercise for particular district types. For example, we could demonstrate the extent to which the outcome of open-seat contests hinged on candidate quality as opposed to the partisanship of the district's voters, employing the "Level Importance" technique described by Achen (1982). [3] Since the CDs are sam-

[1] To illustrate, imagine that our sampled CD involves a highly contested race with a Democratic incumbent in a district where voters are no more Democratic than Republican in their party identification. Let's say that the Democrat wins by a narrow margin in part because of the effect of incumbency—i.e., because voters are more likely to cast their votes for an incumbent than a non-incumbent, *ceteris paribus*. We cannot estimate this effect, since we only observe one race, and hence our explanation of the vote will be incomplete. Still, since there is no variance on this or any other CD-level variable in our study, our estimates of the effects of individual-level variables like party identification are not biased; they are simply not generalizable. For example, the effect of party identification in this district might be weaker than the effect of party identification in races with open seats, and we would not know it.

[2] This is a two-stage cluster sample design. Assuming that each CD has the same number of eligible voters, it generates an equal probability sample of eligible voters in the US.

[3] Sampling at random within a given CD ensures that sample findings can be generalized to the population of eligible voters within that CD. Sampling CDs at random ensures that the sample findings concerning open-seat CDs can be generalized to the

pled at random within the nation, we could draw further conclusions from the data about the overall pattern of congressional election results in the nation as a whole.

These two designs differ in the number of CDs sampled (which we will refer to as "J") and the number of respondents per district (n). The first involves a simple random sample ($n = 3000$, within one CD), whereas the second involves a two-stage, cluster sample; first a sample of CDs (clusters) is selected ($J = 100$), and then individuals are chosen within CDs ($n = 30$). Both designs produce an overall sample (N) of 3000 eligible voters. Neither design involves stratification.

Stratification in the first design would involve sorting the population of eligible voters into categories on one or more stratifying variables before sampling. Within each stratum, respondents would be sampled in proportion to their population frequency if one was striving to achieve an equal probability design. This procedure would ensure that the sample percentage of respondents within a stratum equaled the population percentage, and would increase the power [4] of statistical tests involving dependent variables that were correlated with the stratifying variable(s) (Kish, 1965; Judd et al., 1991). Stratifying also enables one to over-sample within strata—to sample disproportionately so as to increase the representation of a group that otherwise would be represented in the sample in small numbers. In the second design, one could repeat such a procedure at each sampling stage—stratifying the population of CDs before selecting 100 of them at the first stage, and stratifying eligible voters within CDs before selecting 30 of them (within each CD) at the second stage. In important ways that we describe later, stratification at the macro-level can be one of the most crucial aspects of the sample design.

In general, and as these examples have suggested, our ability to use national survey data to understand congressional elections depends on how both districts and individuals are sampled within the study. In order to explicate this point further, and to arrive at general guidelines for sample design, we must consider how the data on individuals and on districts are gathered and analyzed.

---

population of open-seat CDs. Taken together, this means that the whole set of findings can be generalized to the population of eligible voters living within open-seat CDs. If we had random sampling within CDs but a purposive sample of open-seat CDs, we could only generalize results to the population of eligible voters within the sampled CDs. If we had random sampling of CDs but a purposive sample of eligible voters within CDs, then, strictly speaking, we would not be able to generalize any findings concerning individuals at all.

[4] The *power* of a test refers to the probability that one will reject the null hypothesis when the null hypothesis is false (specifically, when a particular alternative hypothesis is true).

3 Multi-Level Data and Analysis

As illustrated by our examples as well as past research (Jacobson, 1997; Brown and Woods, 1992; McPhee and Glaser, 1962)e.g.,, studies of congressional elections are likely to require data on eligible voters, the congressional districts, the campaigns run in those districts, and the candidates contesting the election. This means that the analysis will be based on data characterizing both micro- and macro-level units, integrated into one multi-level dataset. The multi-level structure of the data creates special problems for conventional data analysis.

Below, we briefly discuss the roles played by micro- and macro-level variables in analyses of voting behavior in congressional elections. We then describe different ways of analyzing the integrated, multi-level data: (1) Aggregating all data to the macro-level, (2) treating all data as if it were gathered at the micro-level, and (3) analyzing the micro- and macro-level data simultaneously while also taking into account which data are gathered on which units—i.e., analyzing the data through multi-level modeling. This sets up our subsequent discussion of sample design and statistical efficiency, which assumes that the data will be analyzed through multi-level modeling techniques.

Survey data provide individual-level information on the dependent variable—turnout, vote choice, information about the candidates, and the like. Survey data also provide information on individual characteristics that operate as independent variables, like partisan identification, group characteristics, and issue positions. And survey data provide information on individual characteristics that either mediate or moderate the effects of contextual characteristics on outcomes. For example, a respondent's level of political awareness may influence how he or she responds to the messages of the campaign (Zaller, 1992).

Information on congressional district-level characteristics may be drawn from many sources. Census data, official government records, campaign documents, and mass media sources all contain relevant data. Even survey responses, if aggregated to the CD-level, can provide useful data on candidates or campaigns. One could, for example, use the within-district mean placement of a candidate on a liberal-conservative scale as an index of the candidate's ideology (although for incumbents, of course, other measures of this are readily available, such as indices based on votes cast in Congress). Such a procedure treats survey respondents as informants. [5]

[5] This requires that respondents be randomly sampled within CDs, and for a reliable measure, that the within-district n be relatively large.

The effects of CD-level variables can be thought of in two ways. First, the attributes of districts, candidates, or campaigns can influence vote choices and electoral outcomes. This influence might be either direct, or indirect—i.e., mediated by other CD-level or individual-level characteristics. Second, they can identify contexts which influence how other variables (individual- or CD- level) influence the vote and election outcome. In other words, they can identify contexts across which the explanatory model varies. [6]

There are different ways to represent the relationships between CD-level variables (as independent) and individual-level variables (as both independent and dependent). One way would be to aggregate the individual-level responses to the CD-level, by averaging across individuals within each district, and then to regress an aggregated dependent variable on relevant CD-level variables and on other aggregated individual-level variables. A simple version of this model is depicted below:

$$\overline{Y}_{\cdot j} = \beta_0 + \beta_1 \overline{X}_{\cdot j} + \beta_2 Z_j + E_j \tag{1}$$

In Equation 1, $\overline{Y}_{\cdot j}$ is the mean of the individual-level dependent variable within each CD, $\overline{X}_{\cdot j}$ is the mean of a relevant individual-level variable within each CD, and Z_j is an attribute of each district, measured at the district level (i.e. not using survey data). In such a regression, one would be able to say something about how, say, mean turnout levels vary depending on whether or not an incumbent is running. One problem with this method is that there is no guarantee that the relationship found using aggregated individual-level variables will be the same as the relationship found when using the disaggregated individual-level variables. This is the cross-level inference problem (Achen and Shively, 1995). Related, this model is incapable of representing cross-level interactions, where the effects of micro-level variables vary across macro-level contexts, or vice-versa. Another problem is that, since the typical survey includes many more individuals than districts, the degrees of freedom

[6] If one expects the explanatory model to vary across CDs, then one would build interactions into the model. When those interactions involve an individual-level variable in addition to a CD-level variable they are usually called "cross-level interactions," a term we employ below. Cross-level interactions can be used to represent how attributes of districts shape the influence exerted by some individual-level independent variable (e.g., the notion that in an open-seat contest party identification has more effect the votes than it has in races involving an incumbent), but can also be used to represent how attributes of individuals shape the influence exerted by some district-level independent variable (e.g., the notion that politically unsophisticated voters are more likely to be affected by negative campaigning than politically sophisticated voters). See Fisher (1988) for an excellent discussion of how the same interaction term can be used to estimate coefficients from very different models.

available for hypothesis testing are often drastically reduced via such aggregation. If the survey were using the second hypothetical design we described above, this aggregate regression would have 100 degrees of freedom, despite the availability of information about 3000 individuals.

Alternatively, one might be tempted to model all of the data at the individual level, pretending, in effect, that we have 3000 observations at the CD-level rather than the true number, 100 (again, alluding to the second design example we presented above). Equation 2 presents such a model. In this case, we've added a subscript of i to the CD-level variable Z.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + E_{ij} \tag{2}$$

One problem with this method of modeling the data is that we do not have $N = J \times n$ (number of districts $\times$ number of individuals per district) independent values of Z or of X. Rather, we only observe J independent values of Z, and somewhere between J and N independent values of X. OLS, in this case, would not estimate the correct standard errors.

The number of independent observations obtained in a multi-level design is called the "effective N." In a simple two-stage design like we described in our earlier example, the effective N of macro-level units is J, the number of such units randomly sampled at the first stage. [7] The effective N of micro-level units is more complicated. Because individuals are nested within districts, the values that individual-level variables take on are not likely to be independent within districts. Put another way, we would expect a sample of individuals chosen at random within a district to be more similar to each other than a sample of individuals chosen at random from within the population at large. The most common way to gauge this homogeneity is to summarize it using a statistic known as "rho", the "intraclass correlation coefficient" (ρ). [8] The coefficient ρ ranges from 0 to 1, where $\rho = 0$ corresponds to the case where there is no

[7] If a multi-stage sample design is used where district selection occurs at a later stage, CDs will be clustered within higher-level units and thus the effective N of CDs will be less than the total number represented in the sample. Similarly, as long as the first stage involves sampling areas that contain more than one CD, then even if the CD is not a sampling unit at later stages the effective N of CDs will be less than the number of CDs that fall into the sample. As we describe later, this is true of the sample designs NES has used, with the exceptions of the 1978 and 1980 studies. To estimate the effective N in such cases, one would follow a procedure comparable to the one used for estimating the effective N of micro-level units, which we describe next.

[8] Kish (1965, page 161) introduced this statistic and called it "roh." Most other authors have depicted "rho" by the Greek letter ρ, which is the symbol Kish assigned to this statistic when using it in mathematical formulas. Kish chose "roh" because it is an acronym for "rate of homogeneity."

tendency for individuals nested within macro-level units to be similar to one another, and $\rho = 1$ corresponds to the case where all individuals nested within macro-level units are identical to one another. [9] Kish (1965) used this measure of homogeneity within clusters to calculate the effective N of observations for a given individual-level variable, as depicted below (for the equal cluster size case).

$$\text{effective } N = \frac{N}{1 + (n-1)\rho} = \frac{Jn}{1 + (n-1)\rho} \tag{4}$$

Equation 4 shows that, as the homogeneity within clusters (ρ) increases, then the effective N decreases. [10] Further, when holding the total sample size (N) constant, as the cluster size (n) increases, then the effective N also decreases. Since $N = J \times n$, what this means is that as n increases and J decreases—that is, when the degree of clustering in the design increases—the effective N of micro-level units decreases.

Despite the fact that we can identify the effective N for each of the variables in any given model, the fact that the effective Ns vary across the variables still poses a problem for OLS. [11] The most appropriate model for estimating effects with nested data is the hierarchical, or multi-level model. The multi-level model was developed to represent how the behavior of individuals is influenced by their own (micro-level) characteristics as well as the characteristics of the macro-level contexts in which they are nested (CDs, in our case). It enables one to simultaneously estimate the effects of micro-level variables, macro-level variables, and interaction variables, including cross-level interactions, all while taking into account the multi-level nature of the data. In Appendix A, we provide a very brief description of the model. Jones and Steenbergen (1997) provide a very useful overview, illustrating their discussion with ex-

[9] Equation 3 shows one way to depict the formula for ρ.

$$\rho = \frac{\text{variance between macro-units}}{\text{total variance}} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{3}$$

In this formula, τ^2 represents the between-group variance, and σ^2 represents the within-group variance. As this formula suggests, ρ indicates the proportion of the total variance in some variable that is attributable to the macro-level unit. There are numerous methods of estimating ρ. Later in this paper we use the so-called "ANOVA method" where $\hat{\rho} = \frac{(F-1)J/n}{1+(F-1)J/n}$ (Snijders and Bosker, 1999).

[10] When $\rho = 0$ the effective N is simply N, and when $\rho = 1$ (i.e., all individuals nested within CDs take on the same value) the effective N is J.

[11] OLS is no longer BLUE. What is more, OLS yields biased coefficient estimates if the model estimated involves cross-level interactions (see, for example Kreft and De Leeuw, 1998, especially Chapter 2).

amples from political science. Further information can be found in Bryk and Raudenbush (1992), Goldstein (1999), Kreft and De Leeuw (1998), Longford (1993), Pinheiro and Bates (2000), and Snijders and Bosker (1999).

For our purposes, what is important is how various sample design decisions influence statistical efficiency and the power of hypothesis tests when multi-level models are estimated. We take up this issue next.

4 Sample Design and Efficiency in the Estimation of Macro-level Effects

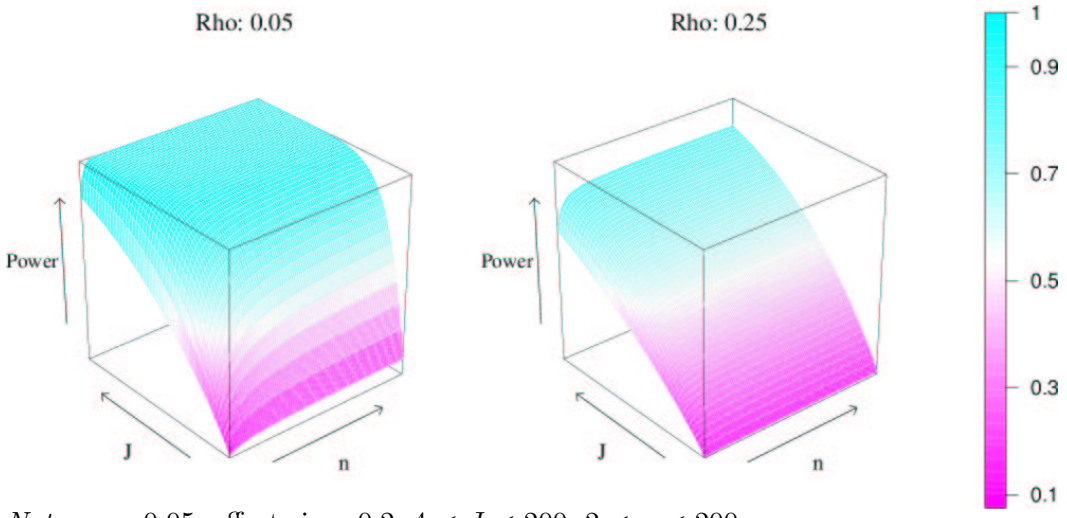
4.1 Number of CDs (J) and Respondents per CD (n)

As our previous discussion has implied, all other things held constant, statistical efficiency in estimating macro-level effects is enhanced by increasing the number of macro-level units—in our case, CDs. This is simply a matter of the effective N of macro-level units. In any analysis involving CD-level explanatory variables, the degrees of freedom available for estimating CD-level effects is determined by taking into account the number of CDs, rather than the number of individuals, that fall into the sample. Hence, the efficiency of statistical estimates (and the power of hypothesis tests) is strongly influenced by the number of CDs sampled.

> 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 this level, are at least as stringent as requirements on the sample size in a single level design with q explanatory variables.(Snijders and Bosker, 1999, page 140)

At the same time, when one is estimating multi-level models, then both the number of macro-level units (J) and the number of micro-level units nested within them (n), and, hence, the total number of micro-level units (N), affects the efficiency of one's estimates, as does ρ (Kreft and De Leeuw, 1998; Mok, 1995; Snijders and Bosker, 1999). [12] To see this, consider Fig. 1, which uses

[12] The estimators of the coefficients in multi-level models are consistent, but not unbiased. As Mok (1995) has demonstrated, increasing J also diminishes the degree of bias in the coefficient estimates. "...consistent with advice given in the classical literature on cluster sampling designs, if resources were available for a sample size



Note: $\alpha = 0.05$, effect size=0.2, $4 < J < 200$, $2 < n < 200$.

Fig. 1. Power of tests for macro-level effects.

an algorithm developed by Raudenbush (1997) to show how the power of statistical tests concerning the effects of CD-level variables is influenced by J, n, and ρ.[13]

In this simulation, we stipulate a very simple model, in which one macro-level variable is seen as affecting a micro-level dependent variable. We assume that the true effect, gauged in terms of a standardized regression coefficient, is 0.1.[14] The two panels show how increasing J and n increase the power of a hypothesis test concerning the effect of the macro-level variable. The left panel portrays the case where $\rho = 0.05$ (i.e. the case where individuals are not very homogeneous within districts) and the right panel portrays the case where $\rho = 0.25$ (i.e. the case where individuals are quite homogeneous within districts).[15] Thus, in this simulation, we are allowing 3 aspects of the design to vary: the number of macro-units (J), the number of micro-units per macro-unit (n), and the amount of homogeneity among micro-units that are nested within macro-level units (ρ). The results in Fig. 1 show that ρ plays an important role in determining the power of tests concerning the macro-level variable; the higher the ρ, the lower the power. Further, increasing both J and

n, comprising J schools with I students from each school, then less bias and more efficiency would be expected from sample designs involving more schools (large J), and fewer students per schools (small I) than sample designs involving fewer schools (small J), and more students per school (large I)" (Mok, 1995, page 6).

[13] Professor Raudenbush sent us the SAS code to implement his algorithm, and we modified it for use in Splus.

[14] The program actually requires us to stipulate an "effect size". We stipulated an effect size of 0.2, which corresponds to a standardized regression (correlation) coefficient of 0.1 (Snijders and Bosker, 1999, page 147).

[15] Estimated values of ρ calculated using NES data are typically in the 0.05 to 0.3 range. See below, Table 6.

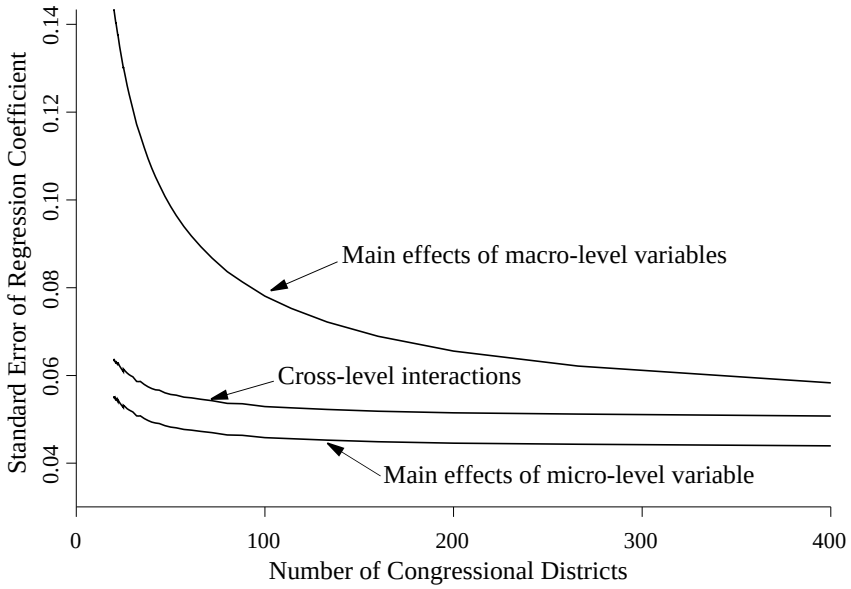
n—and, hence, N—also increases power. More importantly, however, power is much more dramatically enhanced by increasing J than by increasing n. [16]

To further illustrate this tradeoff, we also used a simulation to estimate the standard errors associated with coefficients in a more complex multi-level model. The model called for one micro-level dependent variable to be regressed on (a) one micro-level independent variable, (b) two CD-level independent variables, and (c) two cross-level interactions—i.e., the interactions between each CD-level variable and the micro-level independent variable. For example, one might think of this model as regressing a summary index of knowledge about the candidates on the respondent's level of political awareness, whether or not the election involved an open-seat, whether or not the race was competitive, and the interaction between these CD-level characteristics and the respondent's political awareness level.

This simulation holds the total number of micro-level cases (N) constant, while varying J and n. Hence, as J is increased, n is necessarily decreased, and vice-versa. Fig. 2 shows how the relative sizes of J and n affect the standard errors of the independent variables. [17] In this figure, the x-axis shows the number of congressional districts (J), leaving the number of cases sampled within each district (n) implicit. As J increases, however, n is decreasing. As Fig. 2 shows, increasing J has an enormous effect on the size of the standard errors associated with the macro-level variables. Yet, in this simulation, and as a general matter, the gains (in terms of smaller standard errors) diminish as J increases. The standard errors of the micro-level variables, and of the cross-

[16] As Fig. 1 also shows, the effect on power (in testing for macro-level effects) of increasing the within-district sample size, n, depends on whether ρ is small or large. If ρ is small (.05 in the simulation), then increasing the n is somewhat helpful. If ρ is large (.25 in the simulation), then increasing the n is not helpful. Similarly, a high ρ limits the improvement in power produced by increasing J.

[17] We used the program PINT (Power in Two-Level Designs) developed by Tom Snijders and Roel Bosker (Snijders and Bosker, 1993). We defined the macro-level variables as dichotomous (each scored 0 and 1), with means of 0.5 and variances of 0.25, and as uncorrelated with each other. We defined both independent and dependent micro-level variables as standardized (mean 0 and variance 1), and set the correlation between the macro-level and micro-level explanatory variables to 0. We set $N = 800$, the variance of the residuals at the micro-level to 0.8, and the variance/covariance matrix of the random coefficients (intercept and slope) to $\begin{bmatrix} 0.09 & -0.01 \\ -0.01 & 0.0075 \end{bmatrix}$. Varying these specifications does not alter general conclusions about the tradeoff between J and n as depicted in Fig. 2, but it does matter to the details. In particular, as the variation across electoral contexts in the slope and intercept of the individual-level independent variable increases (i.e., as the main diagonals of the the matrix above increase), increasing J at the expense of n becomes even more desirable.



Note: N is held constant in this simulation, so that as the number of congressional districts increases (J), the within-district n decreases. See text for further details.

Fig. 2. Simulated standard errors of micro-level, macro-level, and cross-level effect estimates.

level interactions, also decrease as J increases, but much less dramatically. This pattern reflects how the effective Ns are changing. For the macro-level variables, the effective N is J. For the micro-level variables and cross-level interactions, the effective N is changing as the clustering (as indexed by n, given a fixed N) in the sample decreases.

The idea that particular characteristics of an electoral context shape how individuals make their voting decisions often seems married to an intuition that one should sample a very large number of individuals within particular electoral contexts. As we have seen, however, this is generally *not* an optimal design strategy. It is much more advantageous to the analyst if J is increased at the expense of n.

If one were especially interested in generalizing about a particular election, then it would, of course, make sense to draw a large and representative sample within the electoral district. That is why, for example, in order to facilitate the study of U.S. Presidential elections, the NES strives to draw a large and nationally representative sample of eligible voters in the U.S. [18] But scholars of congressional elections are not typically interested in explaining the vote in

[18] Even so, such a design is limited in that it lacks variation on explanatory variables that only vary across nations, just as a large-scale study of one congressional district is limited by the absence of variation in explanatory variables that distinguish districts, as we suggested earlier. The Comparative Study of Electoral Systems project (CSES), of which the NES is a part, is a response to this kind of limitation.

a particular district, nor are they likely to believe that the relevant model of the vote (or of other dependent variables) takes a different form in each of the 435 CDs in the U.S. Rather, they are likely to believe that the model varies in systematic and explicit ways depending on the "type" of CD—whether, for example, it involves an open seat or whether the race is highly contested. As such, what is important for analysis is that the total number of individuals falling into each type of CD is not unduly small. [19] Having a large sample size within *each* CD of a given type is not necessary—nor, as we have suggested, is it generally desirable.

4.2 Variances of, and intercorrelations among, CD-level variables

Efficiency in the estimation of causal parameters is also a function of the variance of the explanatory variables as well as the degree of intercorrelation between them (or the linear dependency among a set of three or more). In particular, statistical efficiency in the estimation of CD-level effects is enhanced when the variance on CD-level variables is maximized and the intercorrelation between them is minimized.

These outcomes can be accomplished through stratification in the sampling procedure used to generate survey respondents. In a multi-stage sample design, the probability that any one individual falls into the sample is the product of the probability of selection at each stage. Thus, one can sample disproportionately within strata at the first stage, so as to produce a desired distribution of macro-level units on the stratifying variable, and then compensate at the second stage if one desires an equal probability sample at the micro-level. If, for example, a certain type of CD has a higher probability of selection into the sample than would be warranted by its prevalence in the population, then one would sample individuals within such a CD at a lower rate to compensate.

Both of these points can be clarified through the use of an example. Our example identifies two macro-level stratifying variables, both treated as dichotomous: whether or not the race involves an open seat (*Open*), and whether or not the race is competitive (*Competitive*). Although the general point illustrated by this example does not depend on the particular stratifying variables we have chosen, it is worth briefly addressing why we selected them nonetheless. There are two reasons.

First, they illustrate a general stratification principle: Select stratifying variables that one expects to be important to explaining key dependent variables. The gains from stratification, in terms of statistical efficiency, are a function of

[19] We illustrate this point via simulation in the next section.

Table 1
Population

	Uncompetitive	Competitive
Incumbent	5,000,000 individuals	1,000,000 individuals
	500 CDs @10,000 each	100 CDs @10,000 each
Open-Seat	1,000,000 individuals	1,000,000 individuals
	100 CDs @10,000 each	100 CDs @10,000 each

the extent to which the stratifying variable is related to the dependent variable in question, whether causally or spuriously (Kish, 1965). Research on congressional elections suggests that incumbency and competitiveness—and variables correlated with these—are important to understanding many phenomena of interest. [20] Stratification is helpful whether or not one samples macro-level units within strata in proportion to their population frequency, and whether or not one eventually seeks an equal probability sample of micro-level units. Yet, as we will show, stratifying on key macro-level variables also enables one to employ disproportionate sampling so as to enhance statistical efficiency in estimating macro-level effects. [21]

The second reason is practicality. We have data on these variables for congressional districts in the U.S. from 1948-1998. Later, we generate simulations and present analyses that take advantage of this fact. Further, we have chosen variables easily represented as dichotomies so as to keep the example simple. Thus, we avoided other variables, such as the partisan balance in the district, that might be strong candidates for a stratification scheme.

Suppose, then, that the population distribution of CDs and eligible voters across the four cells defined by these two stratifying variables is as given in Table 1. (We have used unrealistic numbers here—with a total of 8 million people scattered across 800 CDs—to keep things simple.)

[20] The variable *Competitive* is meant to be an indicator of how closely contested the race is, and hence an indicator of campaign intensity(Westlye, 1991). Later in this paper we gauge competitiveness by using information on the electoral margin of victory. Since this kind of information is only available after the election is over, it is not plausible to think of this as information that one could rely upon to make stratification decisions. Margin of victory will, however, be correlated with other CD-level characteristics, like the partisan balance in the district, that one should be able to measure in advance of the election.

[21] If one samples proportionately within strata one can ensure that the percentage of sample units with a given attribute equals the percentage of population units with that attribute. This is an important benefit when random departures from such a result, which are to be expected without stratification, can seriously hamper the analysis. We consider this benefit of stratification in a later section. In this section we focus on the gains from disproportionate stratified sampling.

Table 2
Sample design A, equal probability sample at each stage

	Uncompetitive	Competitive
Incumbent	50 CDs	10 CDs
	$n = 10$	$n = 10$
	$N = 500$	$N = 100$
Open Seat	10 CDs	10 CDs
	$n = 10$	$n = 10$
	$N = 100$	$N = 100$

Table 3
Sample design B, balanced at stage 1, equal probability at stage 2

	Uncompetitive	Competitive
Incumbent	20 CDs	20 CDs
	$n = 25$	$n = 5$
	$N = 500$	$N = 100$
Open Seat	20 CDs	20 CDs
	$n = 5$	$n = 5$
	$N = 100$	$N = 100$

In design A, the population is stratified by the two macro-level variables, *Open* and *Competitive*, CDs are sampled within strata in proportion to their frequency in the population, and then an equal number of individuals is chosen at random within each CD so as to generate an equal probability sample of individuals. This is shown in Table 2. In design B, the population is again stratified by the two CD-level variables, but now CDs are sampled disproportionately in order to create an equal number of CDs in each of the four strata—i.e., a balanced design at the CD-level. Then, an unequal number of individuals is chosen at random within each CD so as to generate an equal probability sample of individuals. This design is shown in Table 3.

Both designs generate equal probability, representative samples of individuals, but they differ in the sample of CDs that is drawn and the pattern of clustering within CDs. In design A, only 20 of the 80 sampled CDs (25%) involve open-seat races, and only 20 (25%) are classified as competitive; taken together, only 10 sampled CDs (13%) involve competitive, open-seat races. In design B, by contrast, the distribution of CDs is balanced on each of the stratifying variables, so that each of the four cells contains 20 CDs (25%). Thinking only of the CD-level aspect of the design, in design A the two stratifying variables have limited variances (.1875 in each case) and a moderate positive correlation (.33). In design B, the two stratifying variables have maximum variances (.25 in each case) and zero intercorrelation.

If all individual-level data were aggregated, so the analysis was performed

at the CD-level (i.e., our sample size in each design is 80), and we simply regressed some aggregated Y variable on *Open* and *Competitive*, design B would be superior in that the two variables are uncorrelated and each have maximum variance, and hence the standard errors associated with estimates of their effects would be smaller in design B than in design A. [22] At the same time, the number of individuals per CD is constant ($n = 10$) in design A but varies in design B ($n = 5$ or 25). This means that design B introduces heteroskedasticity—the mean of an aggregated Y variable would be estimated with varying reliability across the CDs (Hanushek and Jackson, 1977, Chapter 6), and GLS rather than OLS must be used to estimate the model. More importantly, this type of analysis throws away information, as we suggested before, and would only be appropriate if one were interested in cause and effect relationships at the CD-level, whereas most analysts studying subnational elections are interested in explaining individual-level phenomena—or at least first explaining individual-level phenomena and then using those findings to draw implications at the CD- or national-level. This focus requires an analysis which retains individuals as the unit of analysis.

If we work with the data at the individual level, and think of the simplest possible analysis—regressing an individual-level Y on the two contextual variables using OLS—then the two designs are equivalent with respect to a number of things that will influence the estimated standard errors: (1) the number of individuals ($N = 800$), (2) the variances of *Open* and *Competitive* (calculated with the individual-level data), and (3) the correlation between *Open* and *Competitive* (calculated with the individual-level data). In such an analysis, the macro-level variables are treated as attributes of individuals, such that all survey respondents living within, say, a district with an open-seat race, would all be assigned the same value on *Open*. As far as OLS is concerned, there is no difference between these macro-level and other, micro-level variables. Hence, if we were estimating a simple regression model with OLS—which, as we argued earlier, is not advisable—the features distinguishing design A and B would not produce differences in their expected standard errors.

This is useful to notice, we think, but the much more important point concerns how the designs affect statistical efficiency when more appropriate models are estimated—multi-level models which recognize that individuals are nested within macro-level units.

[22] In a trivariate model, the standard error of each slope coefficient is a function of the variance in the stochastic term, the sample size, the variance in the independent variable, and the correlation between the pair of independent variables (Hanushek and Jackson, 1977). The standard error diminishes as the variance in X increases and as the correlation between the independent variables diminishes (approaches zero).

Table 4
Effective sample size

		Design A		Design B	
Actual $N = 800$		Effective $N = 287$		Effective $N = 254$	
500	100	179	36	86	56
100	100	36	36	56	56

One difference between the designs that is taken into account in multi-level modeling involves the effective N that each provides, which varies because of the different clustering entailed in each design. If we assume the intracluster correlation coefficient, ρ, is 0.2, we obtain the effective Ns for each design that are depicted in Table 4. [23]

This analysis suggests that design A is better in terms of the overall effective N, but that design B increases the effective N in the strata that are sparsely populated, and produces something close to parity in the effective sample size across the four cells. [24]

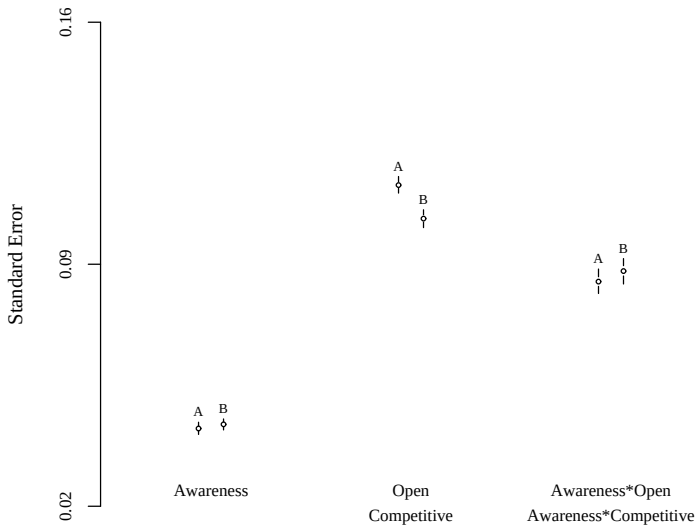
Overall, then, design A is advantaged by its larger effective N, but design B is advantaged in that it maximizes variance in the macro-level variables and minimizes their intercorrelation. As this simple example demonstrates, whether designing a sample that is balanced at the CD-level improves the efficiency of one's estimates depends upon whether the efficiency gain in terms of greater variance in the macro-level variables and lower correlation between them outweighs the efficiency loss incurred by the smaller effective sample size.

To illustrate this tradeoff further, and to add more specificity, we draw on the results of simulations similar to the ones we described earlier, where we generated standard errors associated with the coefficients of a multi-level model that included one micro-level explanatory variable (*Political Awareness*), two macro-level explanatory variables (*Open* and *Competitive*), and the two cross-level interactions. Here, our simulations build in the characteristics of the data implied by design A and design B. [25] Fig. 3 shows the results.

[23] Actually, ρ is likely to be variable across these cells, not constant at 0.2, but we set that complication aside.

[24] By increasing the effective N in the sparsely populated cells—for example, the open seat CDs in our example—design B facilitates subgroup analysis. There are effectively 112 individual-level cases in open seat districts in design B compared to 72 in design A.

[25] Scholars have given close attention to power in two-level designs characterized by equal cluster sizes, but have not given any attention to designs, like our design B, characterized by unequal cluster sizes. Thus, while programs like PINT will estimate standard errors in designs with equal cluster sizes, they do not rec-



Note: Entries depict the standard errors obtained under Designs A and B across 1000 simulations. The dots mark the median SE obtained, while the lines indicate the 5th and 95th percentiles. See the text for further details.

Fig. 3. Simulated standard errors for design A versus design B.

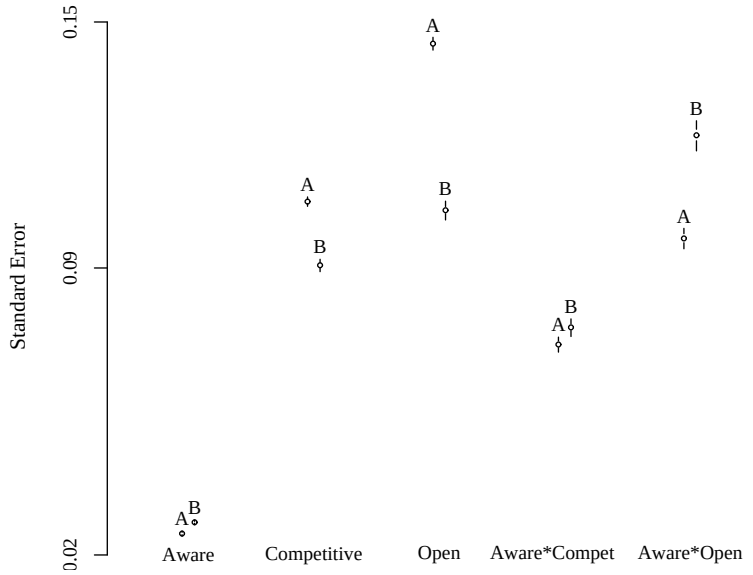
As would be expected, the standard errors on *Awareness* and the cross-level interactions tend to be higher under design B than design A. This reflects design B's smaller effective N. But design B yields more precise estimates of the effects of the macro-level variables. In this case, the benefit from using design B is more substantial than the loss, although the advantage is not overwhelming.

Fig. 4 shows how the standard errors vary across the two designs when the distributions of *Open* and *Competitive* are even more skewed, as is typically the case—with 7% of the races involving open-seats and 13% of the races competitive—and when N and J are set at values more typical of NES studies (1800 cases distributed across 120 CDs). [26] When the population distribution

---

ognize the possibility of unequal cluster sizes. Hence, we developed our own simulations, using the lme package in Splus. For design A, $N = 800$, $J = 80$, $n = 10$, *Open* and *Competitive* each have means of 0.25, variances of 0.1875 and are correlated at 0.33. For design B, $N = 800$, $J = 80$, the ns are either 5 or 25, and *Open* and *Competitive* have means of 0.5, maximum variances (.25) and zero intercorrelation. All other parameters were set to the values given in footnote 17. For each design, we ran 1000 simulations to generate our standard error estimates. The programs which generated the results in Figures 3 and 4, can be found at http://socrates.berkeley.edu/~stoker/sampledesign.html. We are grateful to José Pinheiro for helping us with the lme code.

[26] For design A, $n = 15$ (1800/120). For design B, the average $n = 15$, though the cluster sizes vary across cells of the stratification matrix (from a minimum of 2 to a maximum of 50). For design A, the variance of *Open* is 0.062, the variance



Note: Entries depict the standard errors obtained under Designs A and B across 1000 simulations. The dots mark the median SE obtained, while the lines indicate the 5th and 95th percentiles. See the text for further details.

Fig. 4. Simulated standard errors for design A versus design B: "realistic" parameter values.

of CD-level characteristics is highly skewed, a sampling plan like design A, which simply reproduces that skewed distribution in the sample, can leave analysts with little power for studying CD-level effects. For example, the estimated standard error on *Open* is almost 0.15, which means that a regression coefficient of magnitude 0.25 would fail to achieve statistical significance at conventional levels.[27] Design B, which balances the sample distribution of each CD-level variable, substantially enhances power in this respect. That enhanced power, however, does not come without cost. One pays some price in terms of the efficiency with which micro-level and cross-level effects are estimated.

Although a sample design that is balanced at the macro-level—like design B—can yield more statistical power than one that is not, statistical efficiency is still hampered by one constraint that is common to both designs: the production of an equal probability sample of individuals. This is most easily illustrated by considering the "realistic" version of design B. The constraints built into that design—that $N = 1800$, $J = 120$, the two macro-level variables be uncorrelated and of maximum variance (i.e., that 30 CDs from each stratum

---

of *Competitive* is 0.116, and their intercorrelation is 0.29. Design B again assumes maximum variance in *Open* and *Competitive* (.25) and no intercorrelation. Other parameters were set to the values given in footnote 17.

[27] Since the dependent variables are standardized in these simulations, this means that a true effect size of $\frac{1}{4}$ standard deviation would not be discernable.

are chosen), and the design produce an equal probability sample of micro-level units—together imply that the sample has dramatically different cluster sizes (n) across the CDs falling into the four cells of the stratifying table. In the open/competitive cell, we have very few respondents per CD (60 total respondents across 30 CDs, for an n of 2), but in the incumbent/non-competitive cell, we have many respondents per CD (1500 total respondents across 30 CDs, for an n of 50). Greater efficiency gains could obviously be made if one did not insist on drawing an equal probability sample at the micro-level. One could then select fewer respondents within the incumbent/non-competitive CDs, which would reduce the average cluster size (n), and then increase the overall number of CDs sampled (J). One might also increase the n of respondents selected in open-seat and competitive districts so as to facilitate subgroup analysis. [28]

In sum, sampling so as to produce greater balance in the distribution of macro-level units across the stratifying variables—and thus, maximizing their variances and minimizing their intercorrelation—can substantially enhance statistical power in estimating macro-level effects. This is particularly the case when the stratifying variables identify relatively rare attributes—like the presence of an open seat, or existence of a highly competitive congressional race. One can achieve even greater gains in efficiency if the design need not produce an equal-probability sample at the individual level (which tends to produce large, inefficient clusters); resources that otherwise must be devoted to increasing n can instead be directed toward increasing J.

5 Sample Design and the NES

Although the sample design that the NES has employed over the years has varied, it has typically involved a multi-stage procedure like the one described below for 1988.

> The 1988 NES is based on a multi-stage area probability sample selected from the survey research center's (SRC) [1980] national sample design. Identification of the 1988 NES sample respondents was conducted using a four stage sampling process—a primary stage sampling of U.S. standard metropolitan statistical areas (SMSAs) ... and counties, followed by a second stage sampling of area segments, a third stage sampling of housing units within sampled area segments and concluding with the random selection of

[28] Simulations comparing design A (equal probability at both stages) and design B (balanced at the macro-level) with a design that is balanced at both levels demonstrate that the latter is far superior in terms of statistical efficiency in estimating both macro-level effects and cross-level interaction effects. See http://socrates.berkeley.edu/~stoker/sampledesign.html.

a single respondent from selected housing units. (Miller, 1988)

From the 1988 sampling design emerged 45 1st-stage geographic areas, of which 11 entered at the first stage with a probability of 1 (all large cities), and 34 entered with probability proportionate to their population size. A total of 2040 individuals living in 135 congressional districts responded to the 1988 NES survey, an average of about 15 individuals per district. CDs were partially nested within primary sampling units, so that the effective N of CDs is in the 105-109 range. [29]

Table 5 provides summary information about the sampling design of the NES for each year that congressional district data were gathered since 1956 (a total of 21 years). [30] The only time that the NES departed from its basic area probability sampling design was in 1978 and 1980, when the NES employed a multi-stage procedure that used the CD as the 1st-stage sampling unit. In 1978, 108 CDs were selected at the first stage (J), and roughly 25 individuals per CD (n) were selected in later stages. CDs were first stratified on the basis of a combination of variables that included geographic region, state, urbanization, and recent voting behavior. The individuals sampled within CDs were clustered within lower units, so that the effective N within the districts is substantially lower than 25, on average. As such, the within-district findings can not be generalized to the CD as a whole.

[29] Calculating the effective N requires taking into account the degree of clustering in the sample design, and hence the intraclass correlation, for any given variable of interest. We estimated the intraclass correlation (ρ) for two variables, one characterizing the race as open or as involving an incumbent (*Open*), and one characterizing the margin of victory (*Margin*). The analysis examined the degree to which CDs were clustered within primary sampling units (i.e., what percentage of the total variation in the two variables was between-PSU variation as opposed to within-PSU variation). For *Open* the $\hat{\rho}$ was 0.14 and for *Margin* the $\hat{\rho}$ was 0.12, and the n was, on average, 3. This translates into an effective N of 105 and 109, respectively. In this analysis and in others that follow we relied upon data on U.S. congressional districts over the 1948-1998 period. We started with machine-readable data put together by Gary King, which covered the 1948-1990 period (ICPSR #6311). We extended the dataset through 1998 using data provided to us by Jennifer Steen. We also linked these data to the NES survey data over the period, allowing us to characterize the CDs sampled by the NES and to identify how they differed from the population of CDs. In such analyses, we excluded CDs from Hawaii and Alaska since those states are excluded from the NES sampling frame.

[30] Additional information about the NES study designs is available at http://www.umich.edu/~nes.

Table 5
NES sample designs since 1956

Year	N of CDs	N of respondents	Sample design summary[a]
1956	145	1762	1950 SRC Sampling Frame
			12 sr + 54 nsr = 66 PSUs
1958	141	1450	1950 SRC Sampling Frame
			12 sr + 54 nsr = 66 PSUs
1960	141	1181	1950 SRC Sampling Frame
			12 sr + 54 nsr = 66 PSUs
1964	138	1571	1960 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1966	133	1291	1960 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1968	144	1557	1960 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1970	155	1507	1960 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1972	164	2705	1970 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1974	155	1575	1970 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1976	162	2248	1970 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1978	108	2304	108 CD PSUs
1980	113	1614	108 CD PSUs (1978 Frame)
1982	168	1418	1970 SRC Sampling Frame
			12 sr + 62 nsr = 74 PSUs
1984	134	2257	1980 SRC Sampling Frame
			16 sr + 68 nsr = 84 PSUs (11 + 34 = 45 used)
1986	180	2176	1980 SRC Sampling Frame
			16 sr + 68 nsr = 84 PSUs (16 + 45 = 61 used)
1988	135	2040	1980 SRC Sampling Frame
			16 sr + 68 nsr = 84 PSUs (11 + 34 = 45 used)
1990	121	1980	1980 SRC Sampling Frame
			16 sr + 68 nsr = 84 PSUs (11 + 34 = 45 used)
1992	181	2485	1980 SRC Sampling Frame
			16 sr + 68 nsr = 84 PSUs (16 + 45 = 61 used)
1994	190	1795	1980 SRC Sampling Frame
			16 sr + 68 nsr = 84 PSUs (16 + 45 = 61 used)
1996	246	1714	1990 SRC Sampling Frame
			28 sr + 80 nsr = 108 PSUs (18 + 36 = 44 used)
1998	128	1281	1990 SRC Sampling Frame
			28 sr + 80 nsr = 108 PSUs (18 + 36 = 44 used)

[a] In the "Sample Design Summary" column, we first list the year identifying the SRC sampling frame, and then the number of primary sampling units (PSUs) chosen at the first stage of the sample design. The abbreviation "sr" refers to self-representing (probability of selection=1), while "nsr" refers to non-self representing. When we identify the number used in parenthesis, this means that only a subset of the available PSUs were used in the particular study. A number of studies involved both cross and panel respondents. Only in 1996 did this involve panel respondents who were selected from a different sampling frame (1980) than were the cross-section respondents (1990), and hence from a different collection of PSUs. This is why the number of CDs is so high in 1996.

5.1 J, n, and ρ

Fig. 5 contains boxplots that depict the number of respondents per congressional district (n), and list the number of CDs (J), across the NES studies from 1956-1998. The number of CDs varies over time from a low of 108 (in 1978) to a high of 246 (in 1996), with an average of 150. Most of the over-time variation in the number of CDs falling into the sample is a function of how many many units were selected at the 1st stage of the multi-stage design (the number of primary sampling units, or PSUs; see Table 5). The unusually high number of CDs in 1996—246—occurred because the study included respondents drawn from both the 1990 PSUs and the 1980 PSUs.[31] The number of PSUs represented in the NES data, therefore, was larger in 1996 than in any other NES study over the period.

The within-district sample sizes ranged from a low of 1 to a high of 90. Except for 1978 and 1980, in each year the distribution of cluster sizes tended to be skewed to the left, with a long right tail. Thus the median n tends to be in the range of 6-10, while the means are often substantially larger. Because urban areas fall into the typical NES sample with a high probability at the first stage, most of the CDs with a small n are urban and most with a large n are rural. In other words, NES respondents are more clustered in rural, than in urban CDs.[32]

It is ironic that the smallest J (108) and the largest average n (25) over the time series occurs in 1978, the year NES first used the CD as a primary sampling unit in an effort to advance scholars' ability to study congressional elections. Although there are reasons to seek a large within-district n, as we have suggested—e.g., if one seeks to generalize to a particular district; to produce within-district "climate of opinion" estimates; to use respondents as informants about the district, candidates or campaigns; or to generate reliable aggregate district opinion measures for use in studies of representation—all of these purposes require a large *effective* N of respondents within districts. As mentioned previously, because of the degree of within-district clustering in

[31] Every ten years, after the decennial census, the University of Michigan Survey Research Center redesigns its sampling frame and selects a new set of PSUs. Hence, NES samples are typically drawn from one set of PSUs for four to five election studies, and then when the sampling frame is redesigned, drawn from a new set of PSUs (see Table 5). In the 1996 NES, the panel respondents had originally been selected in 1992 from the 1980 SRC sampling frame, while the fresh cross-section respondents were drawn from the 1990 SRC sampling frame.

[32] Since the 1st stage of the sampling procedure is only redesigned every decade, following the decennial census, there is a very substantial departure from independence in the sampling of CDs and individuals over time. This issue is discussed and illustrated in Appendix B.

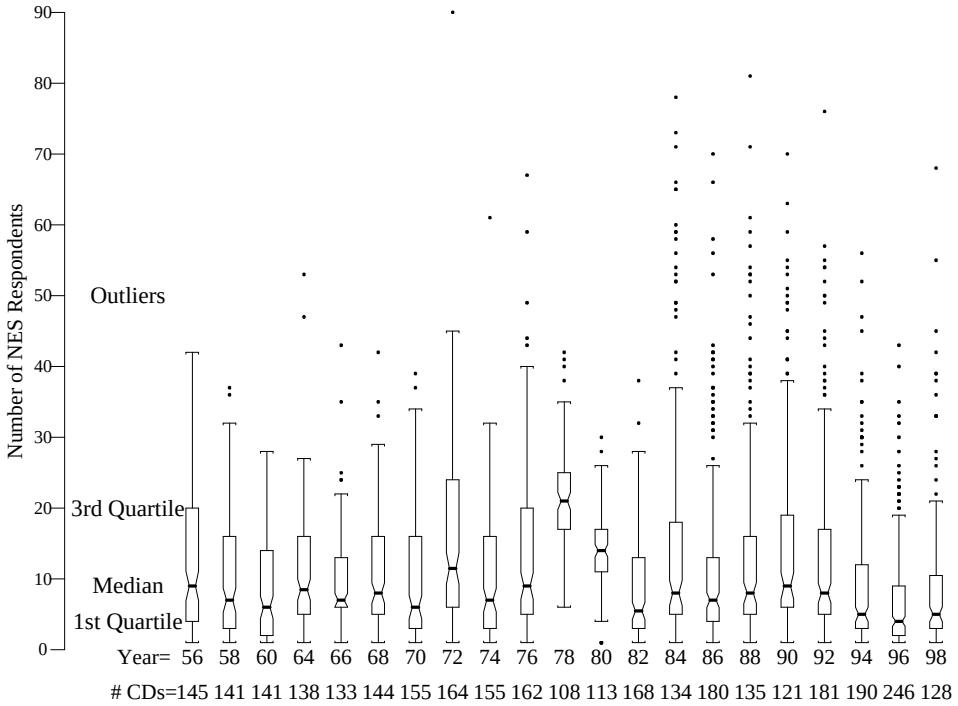


Fig. 5. The number of NES respondents per congressional district, 1956-1998.

the 1978 sample design, this condition is not met. In any event, a large sample of congressional districts still remains very important, in that it substantially affects one's ability to draw conclusions from the data about both the micro- and the macro-level dynamics at work. One could argue, then, that the 1978 design would have been enhanced by increasing J at the expense of n.

As we pointed out earlier, the effective N in a cluster design is a function of ρ(rho) as well as J and n. For micro-level variables, ρ indicates the extent to which respondents are similar to one another within macro-level units. Table 6 shows typical values of $\hat{\rho}$ for an illustrative set of NES variables, based on treating respondents as nested within congressional districts. As would be expected, the $\hat{\rho}$ values are largest when the variables concern CD-specific stimuli. Notice, for example, that the $\hat{\rho}$ is higher for *Incumbent Approval* than for *Congress Approval*. That is, the within-district similarity is higher when respondents are asked about their own district's incumbent than when they are asked about Congress.[33] Notice also that the values of $\hat{\rho}$ are substantially

[33] When one is gathering data on respondents' judgments and choices concerning the candidates in their own districts but striving to analyze the full national sample of respondents, one must construct variables that render the judgments and choices of respondents across districts comparable. This means, for example, recording the respondent's vote choice as a vote for the Democrat or the Republican, even though there is a different pair of Democratic and Republican candidates involved in each district. The ρs will tend to be higher for such variables than for variables where common questions were asked of all respondents concerning common stimuli (e.g.,

higher for two specific variables in the set: *Vote Choice*, and *Incumbent's Perceived Ideology* (average $\hat{\rho}$ values of 0.275 and 0.213, respectively). What such high values of $\hat{\rho}$ reflect is the explanatory importance of CD-level variables. There is more homogeneity within districts on variables like *Vote Choice* and *Incumbent's Perceived Ideology* than on the others in the table; correspondingly, more of the total variation in these two variables is between-district variation.

These high values of $\hat{\rho}$ convey two points simultaneously. First, they remind us, vividly, that CD-level variables are likely to be very important to explaining why voters make the choices they do. Understanding variation in the vote requires understanding how the attributes of candidates, campaigns, and district contexts enter in. Second, they imply that when the design clusters a relatively large number of respondents into a relatively small number of congressional districts, then the effective micro-level N relevant to analyses of the vote will be substantially diminished. In 1988, for example, while the total N was 2040, the effective N given a $\hat{\rho}$ of 0.275 was about 1/5 of that—420.[34] No design modification will influence the fact that, as long as CD-level variables are influential, then individuals who are nested within CDs will tend toward homogeneity. The only response one can make is to try and exploit the features of the design that are under one's control. With ρ reaching magnitudes of 0.2 and even up to 0.5 (see Table 6), the imperative to do so is even stronger. In short, these findings underscore the importance of trying to build more power into one's sample design—by increasing J at the expense of n, and by employing stratification strategically.

5.2 Using CD as a Sampling Unit—Or Not

Most of our earlier discussion has focused on how to design a sample under the assumption that one would use the congressional district as a 1st-stage sampling unit. Yet, with the exception of 1978 and 1980, this has not been the NES practice. There are two important implications that follow from this.

First, if the CD is not a sampling unit, one cannot exert the same kind of control over those aspects of the research design that we have emphasized—

---

evaluations of Congress or of the President); the variance across districts (relative to within) will tend to be higher because the stimulus itself is varying across the districts.

[34] In calculating this figure we used the average *Vote Choice* $\hat{\rho}$ of 0.275, the N of 2040, the average n of 15, and the equation for effective N shown earlier, Equation 4. Even this low estimate of the effective N is probably too large, in that it assumes that the CDs are chosen at random (i.e., that the effective N of CDs is J). But CDs were partially nested within PSUs in 1988.

Table 6. Homogeneity among NES respondents within congressional districts: intraclass correlation coefficients ($\hat{\rho}$)[a]

	Party Identification	Vote Turnout	Vote Choice	Congress Approval	Incumbent Approval	Candidate Recall	Candidate Salience	Incumbent Thermometer	Incumbent's Ideology
Mean $\hat{\rho}$	0.105	0.066	0.275	0.014	0.067	0.102	0.117	0.076	0.213
Minimum $\hat{\rho}$	0.074	0.024	0.111	0.000	0.004	0.063	0.051	0.032	0.122
Maximum $\hat{\rho}$	0.158	0.134	0.495	0.032	0.125	0.207	0.235	0.144	0.293

[a] Entries are summaries (mean, minimum, maximum) of the estimated intraclass correlation coefficient ($\hat{\rho}$) for the variable named in the column across the 21 election studies. $\hat{\rho}$ is calculated based on the decomposition of the total sums of squares for Y (the column variable) into two components: the variation within CDs and the variation between CDs. $\hat{\rho}$ is essentially a ratio of the between-CD variation to the total variation.

J, n, and the characteristics of the CDs sampled (stratification). One can still exert some control, of course, by, for example, increasing the number of PSUs sampled (and hence the number of CDs) and by stratifying PSUs on attributes likely to be correlated with variables important to understanding congressional elections (which is probably not urban/rural). Consider the 1996 NES study, which turned out to have a relatively strong design for the analysis of congressional elections. Because the NES in 1996 reinterviewed some respondents originally selected in 1992 using the 1980 SRC sampling frame, while also interviewing a fresh cross-section of respondents drawn from the 1990 SRC sampling frame, the number of CDs was unusually large and the average n was unusually small. This was probably an unforeseen, but nevertheless fortuitous, side-effect of a design settled upon for reasons that have nothing to do with studying congressional elections. In any event, in terms of the control that can be exerted over the design, one is far better off using the CD as the sampling unit. [35]

The second issue concerns generalization. The NES studies have been designed to generate a representative sample of eligible voters but not a representative sample of congressional districts (except, as noted above, in 1978 and 1980). If the CD is not used as a basis of sampling, then CDs fall into the sample for reasons that are not entirely foreseeable, or at least fall into the sample with no well-defined probability. [36] Since we do not have an equal probability sample of CDs there is no reason to expect the NES sample of CDs to look just like the population. In fact, for example, the NES has tended to overrepresent CDs with Democratic incumbents or where Democrats have tended to win by large margins ($> 10\%$). Correspondingly, the NES has tended to underrepresent CDs with Republican incumbents or large Republican margins of victory. This is demonstrated in Fig. 6. [37]

Without random selection at the CD stage one cannot generalize sample findings about districts to the population of CDs, just as one cannot gener-

[35] It is conceivable that one would field a large enough study to gather survey data on individuals within all 435 districts in the U.S. Then, the CD would not be a sampling unit; instead, the sample design would involve stratification by CD.

[36] In fact, the probability of selection for any given CD could, in principle, be estimated, using information about the probability of selection, and population size, of PSUs and lower-level sampling units. This would not be an easy task—and to our knowledge, it has never been attempted—but if these probabilities were calculated then sampling weights could be devised to remedy the problems concerning generalizability we address here.

[37] We emphasize: This does *not* mean that the NES sample of respondents fails to represent the population from which it is drawn. Except in 1978 and 1980, the NES studies have not been designed to generate a representative sample of CDs, but only to generate a representative sample of eligible voters in the continental United States (excluding Alaska and Hawaii).

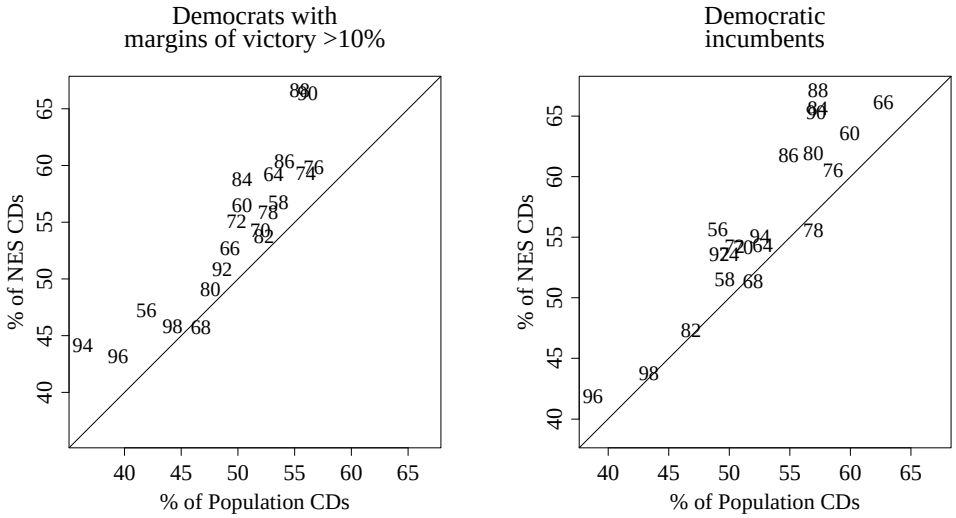
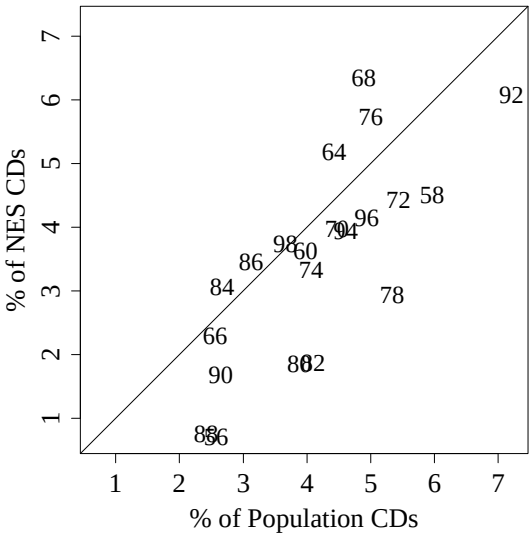


Fig. 6. Percentage of CDs with large democratic margins of victory, and percentage of CDs with democratic incumbents—population percentages vs. NES percentages

alize from a non-probability sample of individuals to a larger population of individuals. In other words, there is no reason to believe that findings based on the CDs that fall into the NES sample would be similar to what one would observe if analyzing the population of CDs as a whole. With a design like the NES, one can only make generalizable statements about individuals, and even then one must avoid statements which sneak in the assumption that the NES sample of CDs is representative. For example, one would be on shaky ground concluding that "Men and women reacted differently to the female congressional candidates running in 1998" if such a conclusion emerged from an analysis of the NES data. There is no particular reason to believe that this same relationship would be observed in an analysis of how voters reacted to the full set of female congressional candidates vying for office across the nation.

5.3 Sampling CDs with Rare Attributes—Or Not

When sampling rare events, one is more likely to undersample them than to oversample them, especially if the number of independent draws is small. With a dichotomous variable, where one outcome is rare (e.g., competitive race) and the other outcome is common (uncompetitive race), this is given by the asymmetry of the binomial distribution. If, for example, 5% of the races in the population are competitive, then it is likely that less than 5% of one's sampled CDs will be competitive. This is one reason why stratification is a useful procedure. If the population were first stratified on the basis of the competitive/uncompetitive variable, then any random deviations from the 5%/95% breakdown in the sample could be avoided. One could, for example, ensure that exactly 5 competitive CDs fell into one's sample of 100 CDs.



Note: We classified a race as competitive if the margin of victory was 10% or less.

Fig. 7. Percentage of CDs with open-seat and competitive races—population percentages vs. NES percentages.

In the case of the NES, the number of independent draws is sufficiently small so as to make this a potential problem, at least with respect to sampling relatively rare events (e.g., CDs with competitive, open-seat races). That number—i.e., the effective N of CDs—is no more than the actual number of CDs that falls into the NES sample, and can be substantially lower than that, given the multi-stage clustered sample design that the NES employs (though will be no smaller than the number of primary sampling units drawn at the first stage of the sample). As mentioned earlier, in 1988, for example, 135 CDs fell into the NES study, but the effective N was in the range of 105 to 109. Fig. 7 shows that, across the 21 election studies we are examining, the NES sample did tend to produce fewer CDs with competitive, open-seat races than were present in the population. This pattern is consistent with the argument we reviewed above, although it is also possible that it is caused by some other aspect of the multi-stage sample design.

The fact that we are vulnerable to undersampling CDs with rare traits is not inconsequential, but the much more serious problem is that, even without such accidents of chance—that is, even if our sample percentage exactly equaled the population percentage—we end up with a sample with very, very few micro- and macro-level cases with the attribute in question. The worst year for the NES in this respect was 1988, where only 22 respondents came from a district with a competitive, open-seat race (1.1% of the total sample), and all of these respondents came from one district (see Table 7). By contrast, 1776 (90.1%) of the 1988 NES respondents, from 121 districts, faced uncompetitive races involving an incumbent. In one of the better years, 1994, the NES sample included 111 respondents (6.7%) from 7 districts in the competitive/open category, and 1254 respondents (75.2%) from 138 districts in the

uncompetitive/incumbent category.

What this means is that the variance in CD-level explanatory variables like *Open* and *Competitive* is typically, and in years like 1988, extremely limited. With so few CD-level cases in three of the four cells of the table obtained by crossing *Open* and *Competitive*, statistical efficiency is seriously diminished. And, if one seeks to elaborate the model by adding additional CD-level variables or by building interactions among them, serious problems involving multicollinearity are likely to arise.

One response in this circumstance, as we suggested earlier, is to alter the sample design by sampling so as to ensure balance in the number of CDs with crucial attributes (our "design B", Table 3). This would essentially even out the CD percentages in Table 9, and generate more statistical power for estimating CD-level effects. This, however, may not be a bold enough step. If the sample design remains an equal probability sample at the individual-level, the alteration just described would leave the percentage of respondents falling into the various CD categories in Table 9 unchanged. Do we really want upwards of 75%—and as high as 90%—of the NES respondents to reside in districts where there is essentially no race?

6 Conclusion

Our advice about how to design a study of congressional elections challenges both conventional wisdom and current practice. If congressional elections are predominantly local affairs, where what is happening in local contexts is critical to understanding the election outcome, then despite intuition to the contrary, we should be designing national surveys that sample relatively few individuals within relatively many CDs. We should also exert control over which CDs fall into our sample, by using CD as a sampling unit and stratifying by important explanatory variables. These stratifying variables are likely to be political, not geographic. Stratification will ensure that our sample contains sufficient variation on key macro-level explanatory variables, enhancing the value of the data to analysts. Even further gains in statistical efficiency can be made by undertaking a national survey that does not generate an equal probability sample of eligible voters (though it should, of course, sample eligible voters with known probability so that sampling weights can be devised and applied). This means oversampling individuals within districts that have rare traits, undersampling in other districts to avoid the inefficiencies of large clusters, and directing resources toward extending the overall CD sample size.

To illustrate and defend these arguments, we have relied on examples, simulations, and data analysis. In this process, we have been quite critical of

Table 7
Congressional districts in the NES: distribution of CDs and respondents[a]

Year		Incumbent Not Competitive	Incumbent Competitive	Open-Seat Not Competitive	Open-Seat Competitive
56	NES CDs	76.1	17.6	5.6	0.7
	NES Rs	76.2	18.2	4.2	1.4
58	NES CDs	66.2	20.3	9.0	4.5
	NES Rs	56.9	30.3	9.1	3.6
60	NES CDs	78.3	14.5	3.6	3.6
	NES Rs	69.9	21.4	5.5	3.1
64	NES CDs	71.1	16.3	7.4	5.2
	NES Rs	66.4	21.9	6.8	4.8
66	NES CDs	78.6	12.2	6.9	2.3
	NES Rs	80.3	11.3	7.2	1.2
68	NES CDs	78.2	9.2	6.3	6.3
	NES Rs	78.0	11.9	4.0	6.1
70	NES CDs	81.5	7.9	6.6	4.0
	NES Rs	84.5	8.2	5.0	2.3
72	NES CDs	75.9	8.9	10.8	4.4
	NES Rs	78.0	7.0	9.0	6.0
74	NES CDs	70.7	16.0	10.0	3.3
	NES Rs	66.8	19.5	7.4	6.2
76	NES CDs	80.3	5.7	8.3	5.7
	NES Rs	78.0	6.4	11.7	3.9
78	NES CDs	80.4	6.9	9.8	2.9
	NES Rs	80.0	7.3	9.2	3.5
80	NES CDs	83.3	11.1	3.7	1.9
	NES Rs	82.4	11.2	4.9	1.5
82	NES CDs	69.4	13.8	15.0	1.9
	NES Rs	71.4	15.0	10.4	3.2
84	NES CDs	83.2	7.6	6.1	3.1
	NES Rs	80.6	10.2	8.5	0.7
86	NES CDs	86.2	4.6	5.7	3.4
	NES Rs	88.4	5.2	4.9	1.4
88	NES CDs	91.7	2.3	5.3	0.8
	NES Rs	90.1	1.2	7.6	1.1
90	NES CDs	88.2	5.0	5.0	1.7
	NES Rs	87.8	3.0	4.6	4.7
92	NES CDs	68.0	13.3	12.7	6.1
	NES Rs	67.4	12.4	15.7	4.5
94	NES CDs	77.5	11.8	6.7	3.9
	NES Rs	75.2	10.0	8.2	6.7
96	NES CDs	75.9	12.9	7.1	4.1
	NES Rs	77.8	12.9	5.5	3.7
98	NES CDs	83.2	9.3	3.7	3.7
	NES Rs	76.8	11.4	1.6	10.2

[a] Cell entries contain %s, describing the distribution of NES cases across the 4 category CD-level variable identified in the top row. The % of NES CDs is the % of CDs in the NES sample (i.e., represented by at least one NES respondent) with the named characteristic. The % of Rs is the % of respondents falling into the named type of CD.

the NES multi-stage cluster sample design. What we have not discussed is that the NES' use of this design reflects cost considerations that flow from the NES commitment to in-person interviewing, which, itself is based on a commitment to data quality. It is simply too costly (in terms of expense) to conduct in-person interviews with a widely dispersed sample of respondents, and too costly (in terms of data quality) to abandon the in-person sampling frame and interview format. Throughout this paper we have not considered the practical difficulties associated with implementing various sample designs, nor situated our advice about sampling within a broader research design framework. But, of course, it makes no sense to encourage researchers to pursue sampling strategies that are too expensive to execute or would compromise data quality. Similarly, while stratification on political variables is desirable from a theoretical standpoint, one needs stratifying variables for which data on population units can be collected, with relative ease, well in advance of the fieldwork. When designing any sample, such considerations must be balanced alongside the kinds of statistical considerations that we have emphasized.

Thus, we think of our advice as opening or reopening, not closing, the discussion about how to design surveys of congressional elections. At the same time, although we have mentioned "congressional" 53 times, "district" 115 times and "NES" 116 times so far, this paper has not just been about designing research focused on congressional elections in the United States. Our arguments are applicable to any research problem involving multiple levels of analysis. The most obvious extension is to research focused on subnational elections in other countries. But multi-level problems are many and varied. Researchers studying political institutions examine bureaucrats nested within bureaucracies and legislators nested within legislatures. Political communications researchers gather data on newspaper articles nested within newspapers and advertisements nested within campaigns. In each of these cases and others, researchers must decide how to trade J for n, how to stratify, and must grapple with ρ. We hope that this discussion, while focused on the study of congressional elections in the U.S., sheds light on the design issues scholars facing other multi-level problems must confront.

Acknowledgments

In preparing this paper, we benefited from helpful conversations with Charles Franklin, Don Green, Jim Lepkowski, Sam Lucas, Richard Gonzales, José Pinheiro, Stephen Raudenbush, Tom Snijders, and Charles Stein. Ben Highton, Mark Franklin, Tom Piazza, and Chris Wlezien gave us helpful comments on an earlier draft. The Institute of Governmental Studies and Survey Research Center at UC-Berkeley provided valuable research support. Our analysis uses the program PINT (Power in Two-Level Designs) developed by Tom Snijders

and Roel Bosker, available at http://stat.gamma.rug.nl/snijders/. We are also grateful to Stephen Raudenbush for providing us with a beta-version of Optimal Design, and the associated algorithm for calculating power in multi-level designs, and to Liu Xiaofeng for helping us with that transaction. Finally, we thank Jennifer Steen for providing us with machine-readable data on U.S. congressional districts from 1992-1998, and Pat Luevano of the NES staff for helping us sort out various discrepancies between the district-level data and the NES data. We alone retain responsibility for any errors of fact or interpretation.

Appendix A. The Multi-Level Model

The multi-level model was developed to represent how the behavior of individuals is influenced by their own (individual-level) characteristics as well as the characteristics of the contexts in which they are nested (CDs, in our case). Below, we illustrate the model for the case in which there are two district-level explanatory variables and one individual-level explanatory variable. Interested readers should consult Bryk and Raudenbush (1992), Goldstein (1999), Jones and Steenbergen (1997), Kreft and De Leeuw (1998), and Snijders and Bosker (1999) for further details.

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \tag{A1}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}z_{1j} + \gamma_{02}z_{2j} + u_{0j} \tag{A2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}z_{1j} + \gamma_{12}z_{2j} + u_{1j} \tag{A3}$$

Combining the previous three equations, we have:

$$y_{ij} = \gamma_{00} + \gamma_{01}z_{1j} + \gamma_{02}z_{2j} + u_{0j} + (x_{ij})(\gamma_{10} + \gamma_{11}z_{1j} + \gamma_{12}z_{2j} + u_{1j}) + e_{ij} \tag{A4}$$

$$= \gamma_{00} + \gamma_{01}z_{1j} + \gamma_{02}z_{2j} + \gamma_{10}x_{ij} + \gamma_{11}z_{1j}x_{ij} + \gamma_{12}z_{2j}x_{ij} + (u_{0j} + u_{1j}x_{ij} + e_{ij}), \tag{A5}$$

where $j = 1 \ldots J$ and $i = 1 \ldots n_j$.

Equation A1 specifies the relationship at the micro level, (where x_{ij} is the individual-level independent variable). Equations A2 and A3 specify the relationship between the coefficients of the micro-level equation and the macro-level variables. Finally, equations A4 and A5 combine the previous equations into a single equation. With this model, one can evaluate the extent to which a macro-level variable (z_{1j} or z_{2j}) directly influences the dependent variable. At the same time, the micro-level regression coefficients β, are allowed to vary across macro-level units. That is, each macro-level unit (j) is allowed to have its own intercept (β_{0j}) and slope (β_{1j}). The mean of the distributions of these

intercepts and slopes is summarized by the γ terms found in equation A5. The term γ_{00} is the overall mean of y_{ij}; γ_{01} and γ_{02} indicate the (average) direct effects of the macro-level variables on y_{ij}; γ_{10} indicates the (average) direct effect of the micro-level variable; γ_{11} and γ_{12} indicate the (average) effect of the cross-level interactions.

In equation A5 the terms from equation A4 have been reordered to put all of the error terms at the end, in parentheses. Notice that this error component includes a term that refers to the micro-level independent variable. The existence of this term is taken into account in multi-level model estimation, but would not be taken into account if one simply tried to estimate the effect of micro-level, macro-level, and cross-level interaction variables with OLS (hence, resulting in bias with OLS estimation). This error component also includes terms (u_{0j} and u_{1j}) that refer to the variance in the macro-level intercepts and slopes, respectively. And, it contains the micro-level disturbance (e_{ij}). Either Maximum Likelihood or Generalized Least Squares is necessary to efficiently estimate the model (Snijders and Bosker, 1999, pages 56-57).

Appendix B. Non-independence across time in the NES sample of CDs

The first stage of the NES sampling procedure, involving the selection of PSUs, is revised every ten years, in the wake of the decennial census. Because of this, once a CD falls into the sample it tends to stay there for study after study, until the next census is taken and the NES sampling frame is revised. In other words, there is a substantial departure from independence in the sampling of CDs (and individuals) over time. The extent of this departure is illustrated in Figure 8. Along the X-axis is the number of times that a CD fell into the NES sample over the 21 election studies between 1956 and 1998 where CD information is available from NES. The height of the bars represents the proportion of CDs falling into each category along the X-axis. Thus, for example, the first bar shows that 14.8% of the CDs were never represented in any NES sample over the period, while the last bar shows that 0.6% were represented 20 out of 21 times. The area under the illustrated density curve indicates the expected percentages under an assumption of independence of the draws. [38] The contrast is stark.

This figure can be viewed as illustrative, but no more, in that it builds in a problematic assumption—namely, that the identity of a given CD is unchanging over the entire time span. In developing this figure, we first threw out all

[38] The figure depicts the binomial density with P set to 0.35, since on average over the period the NES sample included 35% of the population of CDs.

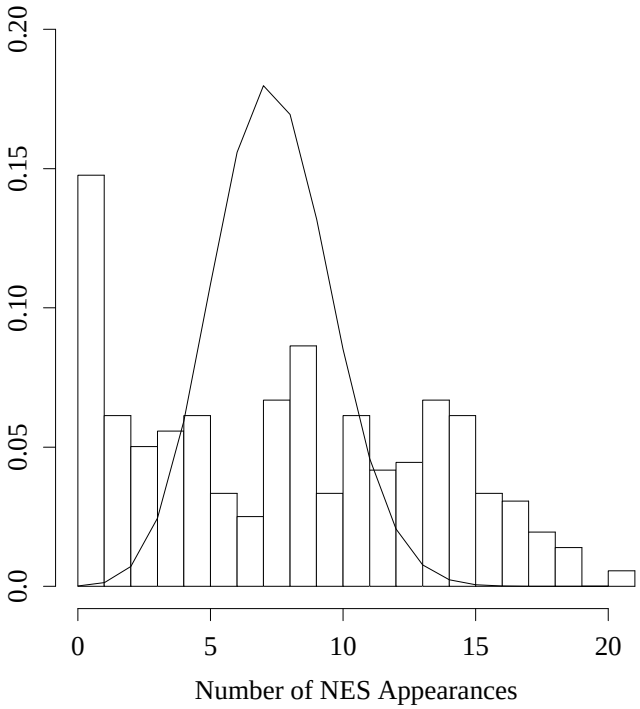


Fig. 8. Empirical distribution of CD appearances in the NES, and expected distribution under an assumption of independence in the draws.

CDs that did not "endure" over the 1956-1998 period. Some were created after 1956, as a function of redistricting, and some that once existed later disappeared, again because of redistricting. By our count, a total of 512 CDs were in existence for at least one election over the 32 years; 359 were in existence for the entire period. This assumes, however, that any district which retained its unique identifying number over time was actually the same unit at each point in time. This, however, is clearly wrong. Minnesota District 2, for example, may have completely been transformed in its boundaries, perhaps several times, over the period. The over-time dependence in the sample of CDs, illustrated in Figure 8, only becomes an issue for research that analyzes CD-level data over time. Yet such research cannot even be begun without solving the seemingly intractable "unit change" problem just described.

References

Achen, C. H., 1982. Interpreting and Using Regression. Sage, Newbury Park, CA.

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

Brown, R. D., Woods, J. A., 1992. Toward a model of congressional elections. Journal of Politics 53 (2), 454–473.

Bryk, A. S., Raudenbush, S. W., 1992. Hierarchical Linear Models. Sage Publications, Newbury Park, CA.

Fisher, G. A., 1988. Problems in the use and interpretation of product variables. In: Long, J. S. (Ed.), Common Problems, Proper Solutions: Avoiding Error in Quantitative Research. Sage, Newbury Park, CA, pp. 86–107.

Goldstein, H., 1999. Multilevel Statistical Models. Edward Arnold, London, internet Edition, available at http://www.arnoldpublishers.com/support/goldstein.htm.

Hanushek, E. A., Jackson, J. E., 1977. Statistical Methods for Social Scientists. Academic Press, Inc., San Diego, CA.

Jacobson, G. C., 1997. The Politics of Congressional Elections, 4th Edition. Longman.

Jones, B. S., Steenbergen, M. R., 1997. Modeling multilevel data structures, paper prepared for the 14th annual meeting of the Political Methodology Society, Columbus OH, July 25, 1997.

Judd, C. M., Smith, E. R., Kidder, L. H., 1991. Research Methods in Social Relations, 6th Edition. Harcourt Brace Jovanovich.

Kish, L., 1965. Survey Sampling. John Wiley and Sons, New York, NY.

Kreft, I., De Leeuw, J., 1998. Introducing Multilevel Modeling. Sage, London, UK.

Longford, N. T., 1993. Random Coefficient Models. Clarendon Press, Oxford.

McPhee, W. N., Glaser, W. A., 1962. Public Opinion and Congressional Elections. Free Press, New York, editors.

Miller, W. E., 1988. Codebook: American National Election Studies:1988 Pre-Post Study. Center for Political Studies, University of Michigan.

Miller, W. E., Stokes, D. E., 1963. Constituency influence in congress. American Political Science Review 57, 45–56.

Mok, M., 1995. Sample size requirements for 2-level designs in educational research, unpublished manuscript.

Pinheiro, J. C., Bates, D. M., 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York.

Raudenbush, S. W., 1997. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods 2 (2), 173–185.

Shaw, D. R., 1999a. The effect of tv ads and candidate appearances on statewide presidential votes, 1988-96. American Political Science Review 93 (2), 345–361.

Shaw, D. R., 1999b. The methods behind the madness: Presidential electoral college strategies, 1988-1996. Journal of Politics 61 (4), 893–913.

Snijders, T., Bosker, R., Fall 1993. Standard errors and sample sizes for two-level research. Journal of Educational Statistics 18 (3), 237–259.

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

Westlye, M. C., 1991. Senate Elections and Campaign Intensity. Johns Hopkins University Press, Baltimore.

Zaller, J. R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, New York.