Randomization Tests and Multi-Level Data in U.S. State Politics

Robert S. Erikson, Pablo M. Pinto and Kelly T. Rader

State Politics & Policy Quarterly 2010 10: 180
DOI: 10.1177/153244001001000204

Randomization Tests and Multi-Level Data in U.S. State Politics

Robert S. Erikson, *Columbia University*
Pablo M. Pinto, *Columbia University*
Kelly T. Rader, *Columbia University*

ABSTRACT

Many hypotheses in U.S. state politics research are multi-level, positing that state-level variables affect individual-level behavior. Unadjusted standard errors for state-level variables are too small, leading to overconfidence and possible false rejection of null hypotheses. Primo, Jacobsmeier, and Milyo (2007) explore this problem in their reanalysis of Wolfinger, Highton, and Mullin's (2005) data on the effects of post-registration laws on voter turnout. Primo et al. advocate the use of clustered standard errors to solve the overconfidence problem, but we offer an alternative solution: randomization tests. Randomization tests are non-parametric tests that do not rely on comparisons to theoretical test statistic distributions. Instead, they use distributions tailored to the data, created by randomly scrambling the data many times to simulate what would be observed under the null hypothesis. Unlike with clustering, with the randomization test, U.S. state-level reforms generally fail to be significant both as additive effects and as interactions with individual characteristics.

DO VARIOUS U.S. STATE-LEVEL reforms induce greater turnout at the polls in the U.S.? While some literature on this topic is time-serial in design, much of the reported evidence is cross-sectional in nature. A decided limitation for cross-sectional analysis would seem to be the limiting *N* of 50 cases when the states are the units of analysis. Cross-sectional analysis seemingly undergoes a great improvement when the unit of analysis shifts to the individual voter. Beginning with Wolfinger and Rosenstone (1980), many studies exploit the U.S. Census's biennial post-election survey of voter participation, which includes multiple thousands of respondents. With more than 40,000 cases, rather than a mere 50, the advantage of using Current Population Survey (CPS) data would seem considerable. Not only does one gain in the number of cases but also in the efficiency for estimating effects at the state level, while controlling for effects at the individual level that would otherwise add noise to the model.

State Politics and Policy Quarterly, Vol. 10, No. 2 (Summer 2010): pp. 180–198

One recent example is Wolfinger, Highton, and Mullin's (2005) analysis (hereafter WHM) of the effects of state laws, which alter the costs of voting, on the probability that a registered voter will turn out to vote. According to this analysis, reforms—such as early voting hours, mailed sample ballots, and many other innovations—serve to boost turnout among already registered voters. As we will see, however, the claim is in dispute because of the multi-level nature of the problem. The independent variables of interest, or treatments, are administered at the state level, while the responses, or outcomes, are by individual potential voters.

A troublesome aspect of this multi-level design is the accurate portrayal of the standard errors of the state-level law effects. When they are estimated in the usual way, the standard errors are deflated (and the significance *inflated*), in that the effective *N* for the state effects is treated as the number of individual cases instead of the number of states. The classic statement is by Moulton (1986, 1990) (see also Donald and Lang 2007). Arceneaux and Nickerson (2009) present the general argument regarding how this problem applies to multi-level data.

Primo, Jacobsmeier, and Milyo (2007) (hereafter PJM) offer a useful illustration when applied to WHM's analysis of state laws affecting voting among registrants. PJM show that when the standard errors for the state laws are correctly estimated as clustered standard errors (clustered by states), the coefficients for the various state laws are mainly outside the range of statistical significance. Thus, it would seem that the evidence remains rather unsettled as to whether reform legislation actually boosts turnout among the registered.

At this point, we might conclude that the lesson is complete. Cluster the standard errors, and the standard errors for state-level variables revert to their proper values, where the degrees of freedom are based on the number of states rather than on the number of respondents. Nonetheless, while clustered standard errors are certainly an improvement over naive standard errors, even clustered standard errors can be overconfident for some types of data. The cluster-robust standard error algorithm was designed for data in which the number of groups is large relative to the number of observations in groups. Fifty clusters is generally considered enough for the asymptotic properties of clustered standard errors to kick in, but currently no consensus exists on how many clusters is too few for inference (Angrist and Pischke 2009). WHM analyze 42 states. Is this close enough?

An additional troublesome detail remains in PJM's presentation. WHM's original model includes interaction effects between state-level reforms and individual-level characteristics, which are plausible because the influence of certain reforms will vary by the type of registrant. For instance, notification of one's polling place could be most effective for registrants with little education. As we will see, certain multi-level interactions remain significant with clustering, even when their main effects are decidedly nonsignificant. This could result simply because each of the postregistration laws WHM examines acts only on certain subgroups of registrants. Alternatively, this pattern of significant interaction effects and imprecisely-measured main effects could be a symptom that clustered standard errors are overconfident for multi-level interaction effects. Indeed, recent Monte Carlo evidence suggests that clustered standard errors on multi-level interaction effects are too small, even when the number of groups is large (Leoni 2009).

The present study offers a further insight into the multi-level inference issue by applying a "randomization" test. Through our randomization test, the codes for the 42 states in the analysis are shuffled to provide randomly generated state identifications; each of these identifications is associated with different electoral laws. This test acts as if the state labels are dropped on the floor and then reinserted randomly. Each state has a state code, but it has only a one-in-42 chance of being assigned the correct one. For example, Arizona respondents might be assigned Pennsylvania's laws. The random shuffle is repeated 1,000 times.

The data from this exercise provide an approximation of what to expect if the null hypotheses of no state law effects were true or, in other words, the "placebo law effect" (see Helland and Tabarrok 2004). For each coefficient, we obtain an empirically observed standard error from the standard deviation of the density of coefficients estimated from the 1,000 fake datasets, centered (theoretically) on a value of zero. By this experiment, we randomize knowing the "correct" answer (no state effects) and see whether the actual real-world results are within the 0.05 bounds of significance from our empirical distribution of placebo effects. This test is particularly useful where the theoretical standard error is not easily derived.

Because it is empirically derived, the randomization test properly accounts for any issues arising from clustering, multi-level interaction, or non-normalities in the data. Thus, we can use the randomization test results as both a method for correctly calculating standard errors for a given dataset without resorting to parametric assumptions and as a robustness check against parametric methods like clustering. If the randomization standard errors on the state-level terms and multi-level interaction terms are similar in size to the clustered standard errors, then our intuition that the clustering algorithm is lacking in this setting will be wrong. If, however, the randomization standard errors are larger than the clustered standard errors, and their associated p-values are larger, then we can conclude that the WHM data are not ideally suited for the parametric clustering fix. In general, parametric techniques are valid only if they generate *p*-values close in size to those from a randomization test (Moore et al. 2003, 57; Edgington and Onghena 2007, 289).

In the following analysis, our randomization tests show that clustered standard errors, while a great improvement over typical standard errors, are indeed too small given the structure of the WHM data. In this case, randomization tests are more appropriate for statistical inference. While we use the WHM data to demonstrate the randomization technique, the puzzle of estimating contextual effects that interact with individual characteristics is, of course, not limited to this study. Some recent examples include Griffin and Keane (2006) on African-American turnout; Hogan (2005) on coattails in state legislative elections; Lawless (2004) on female representation in Congress; and Soss, Langbein, and Metelko (2003) on white opinion on the death penalty. Furthermore, our findings that clustered standard errors can be overconfident for multi-level interaction effects comport with Monte Carlo analysis (e.g., Leoni 2009). This suggests that the randomization test approach should be more widely applied.

THE RESEARCH QUESTION

WHM (2005) use individual-level data from the 2000 Voter Supplement of the Current Population Survey to test the effects of U.S. state post-registration laws on the likelihood of turnout among individuals registered to vote. While turnout among registered individuals is already relatively high, averaging 86 percent in their sample, WHM hypothesize that certain state laws further decrease the cost of voting, even for those who are already registered. Extended voting hours in the morning and evening, time off from work on Election Day for public and private employees, and receipt of sample ballots and polling place information in the mail give potential voters more time and more information. These laws, then, should be associated with higher voter turnout.

WHM also hypothesize that the effects of certain post-registration laws should vary across different subgroups. For example, time off work for public employees should primarily affect the likelihood that public employees will vote, as opposed to private employees or the unemployed. Receiving sample ballots in the mail should primarily affect the voting likelihood of individuals who do not already have the information, like young and less educated people. To accommodate these potential across-group differences, WHM set up a model with several interaction effects between individual characteristics and state laws. We replicate their findings in Table 1 below.[1]

We do not object to the theory behind WHM's hypotheses about the effects of U.S. state post-registration laws on individual turnout.[2] Like PJM, however, we have concerns about the precision with which those effects can be estimated given the multi-level structure of the data. As PJM argue, the WHM data is generated by a process that includes a compound error term. One part of the error is at the individual level, and one part is at the state level. The state-level component induces clustering among respondents in the same state. Standard regression techniques, like that employed by WHM, ignore the state-level error component, and therefore, they overstate the confidence with which they estimate the effects of state-level variables. We have an additional concern, not addressed in PJM, that standard techniques also overstate the confidence with which they estimate state-level/individual-level interactions.

There are several ways to deal with this compound error term. Strategies include using clustered standard errors, modeling state random effects, employing a full hierarchical linear model, and using OLS on data aggregated to the state level. Arceneaux and Nickerson (2009) show that under the "ideal conditions" of random treatment assignment and normally-distributed cluster-level and individual-level disturbances, each of these techniques performs equally well. However, in the case of observational data, such ideal conditions are rarely met, and estimated standard errors might vary across models because of the different assumptions imposed by each technique. For the WHM data, PJM advocate the use of clustered standard errors over hierarchical linear modeling because, theoretically, clustering makes fewer assumptions, and practically, clustering is easier to implement with available software. Nonetheless, Leoni (2009) finds that clustering, particularly for data sets in which the number of groups is small relative to the number of observations within groups, yields standard errors that are overconfident, especially on multi-level interaction terms.

We argue that a randomization test is appropriate for assessing multi-level hypotheses like those in WHM, particularly those with multi-level interaction effects because, unlike the parametric methods discussed above, randomization tests do not rely on any distributional assumptions about the disturbances in the model.

THE RANDOMIZATION DESIGN

Randomization or permutation tests are a non-parametric way to derive standard errors and significance tests for the effect of a variable on an outcome. They are used widely in biology (e.g. Manly 1997) and increasingly in economics and business applications (e.g. Kennedy and Cade 1996).

Typically, we want to determine the likelihood that an estimated coefficient is different from the null hypothesis, usually zero. Standard parametric methods use some function of the coefficient and the estimated standard errors to calculate a test statistic that is theoretically distributed in some way and compare that test statistic to its reference distribution. If that test statistic is relatively rare, we can be confident that the estimated coefficient is different from the null hypothesis. For example, in the simple case of OLS, we estimate standard errors with the assumption that disturbances are distributed i.i.d. $N(0, \sigma^2)$. We derive a *t*-test statistic by taking the ratio of the estimated coefficient and the estimated standard error and compare that statistic to a student's *t* distribution. If the test statistic is larger than the critical value 1.96 or smaller than -1.96 (for sample sizes 1,000 or larger), we reject the null hypothesis of no effect at the 95 percent confidence level.

Randomization tests proceed in an analogous way but without relying on theoretical distributions. First, we estimate the coefficient on a variable of interest and its associated test statistic using our preferred model specification. Then, we randomly reshuffle the data in such a way that no systematic relationship occurs between the variable and the observed outcome. Then, we rerun our preferred model on the shuffled data and obtain a new estimate of the coefficient and test statistic. We reshuffle and re-estimate 1,000 times. This process provides a distribution of 1,000 estimated coefficients and 1,000 test statistics centered, theoretically, at zero. This is the reference distribution for the randomization test. By locating the observed effect (the estimated test statistic) on this distribution, we are able to assess the probability that the effect could have occurred by chance.

Randomization tests were originally developed by Fisher (1935) to test the effect of a treatment in a randomized experiment. Because experiments typically have smaller sample sizes, it is possible to shuffle the data to represent all of the possible permutations of treatment to subject. Then, the randomization test will be exact. However, in many observational contexts, obtaining an exact randomization test is practically infeasible. Sampling many times from the set of possible permutations, however, provides an approximate randomization test.[3] Manly (1997) argues that, for 95 percent confidence levels, randomization tests using 1,000 draws should be powerful enough to detect an effect.

Unlike parametric significance tests, randomization tests make no assumptions about the distribution of disturbances in a model and do not require the distribution of the test statistics to be known. Thus, randomization tests are particularly useful in models with complicated error structures for which theoretical distributions of the standard errors are difficult to derive.

Nevertheless, randomization tests do make one important assumption about the disturbances: that they are exchangeable. Exchangeability means that if the null hypothesis is true, if the variable of interest indeed has no effect, then observed outcomes across individuals would be similar (conditional on confounding covariates) no matter the level of the variable of interest. In other words, if exchangeability holds, then under the null hypothesis, the variable of interest is merely a label that can be applied to any observation without changing the expected outcome. This justifies the shuffling procedure. Exchangeability is a weaker condition than the standard i.i.d. assumption, or in the case of clustering, that observations are independent across clusters, since i.i.d. implies exchangeability but not vice versa.

Widespread agreement exists about how to conduct a randomization test using data that were generated by a randomized experiment (for a review relevant to political scientists, see Keele, McConnaughy, and White 2008) and about how to test the significance of a coefficient in a univariate model (e.g., Manly 1997). However, in the case of observational data and multivariate models, many methods have been devised, and only recently have they been subjected to side-by-side comparisons.

There are several proposed ways to reshuffle data to break the relationship between a variable of interest *Z* and an outcome *Y* for a multivariate model with other covariates *X*. Kennedy (1995) reviews the most common methods in the literature. These include simply shuffling *Z* or shuffling *Y*. More complicated methods include "residualizing *Y*"—residualizing Y with respect to the other covariates *X*, shuffling the residualized *Y*, and regressing it on *Z*—and "shuffling residuals"—regressing *Y* on *X*, shuffling the residuals from this regression, adding them to the predicted *Y* from this regression, and regressing the new *Y* vector on *X* and *Z*.

The results from Monte Carlo analyses in Kennedy and Cade (1996) suggest that the simple method of shuffling *Z* is sufficient in the multivariate context so long as inferences are based on the distribution of test statistics and not on the distribution of coefficients.[4] Further Monte Carlo analyses in O'Gorman (2005) confirm that the simple shuffle *Z* method performs as well in terms of power and size, even in the presence of non-normal error structures and high correlation between *Z* and the other covariates *X*. Thus, we chose to use the shuffle *Z* method: to break the relationship between state laws and individual turnout, we randomly reassign the laws of one state to the residents of another state.

This approach has precedents in empirical social science work. Used as it is here to evaluate the effects of state laws on individual behavior, the shuffle *Z* method is equivalent to the "placebo laws" technique used in both Helland and Tabarrok (2004), to test the effect of "shall issue" gun laws on crime, and in Donohue and Wolfers (2006), to test the deterrent effect of capital punishment.

FINDINGS

To reassess the findings in WHM, we performed a randomization test to estimate empirically derived standard errors and to test statistics for the coefficients in their original model. First, we detached the state-level variables from the individual-level observations. Then, we randomly reassigned state level-variables to state populations. This particular randomization procedure preserves the menu of state laws and the association of individuals within a state. For example, all of the residents of Georgia might be randomly assigned all of the post-registration laws from Washington. It is possible that the residents of Georgia might be randomly assigned the laws of Georgia and even that we could recreate the actual dataset through random assignment. Nonetheless, this is justified because, under the null hypothesis, the actual data is considered to be equally likely as any other permutation.

After shuffling the state laws, we recalculate the interaction terms between state-level and individual-level effects. Finally, we rerun the WHM model and collect the coefficients and z-values from standard significance tests. We repeat this process 1,000 times.

Table 1 replicates the multivariate logit model of turnout in WHM, with unadjusted and clustered standard errors, and shows the estimated standard errors from our randomization method. The logit analysis shows coefficients for six main effects of state laws and six additional interactions involving individual characteristics. Focusing first on WHM's unadjusted standard errors, we see nine of the 12 coefficients pass the conventional 0.05 threshold of statistical significance.

PJM's clustered standard errors offer a useful correction. Of the six additive effects of state laws, only one remains statistically significant and that is time off for private employees, which has the incorrect sign. With clustering, their standard errors expand so that the observed coefficients (with the one odd exception) do not reach the requirement for significance at the conventional 0.05 level. But the six interaction effects show little change in their standard errors from clustering by states. The three interactions that are significant with unadjusted standard errors remain significant (at least at 0.10) with clustering. Perhaps this result means that, as suggested by the Monte Carlo analysis in Leoni (2009), clustering does not properly calculate standard errors on multi-level interaction effects.

Standard errors based on our randomization tests offer the corrective.

Table 1. Comparison of Standard Errors in Full Turnout Model

	Coefficient	Unadjusted Standard Errors	Clustered Standard Errors	Randomized Standard Errors
Early voting	0.14	0.03**	0.10	0.12
Late voting	0.08	0.04*	0.08	0.11
Mailed polling place information	0.24	0.12*	0.22	0.25
Mailed polling place information × education	–0.08	0.04*	0.04#	0.07
Mailed sample ballots	0.29	0.12*	0.18	0.27
Mailed sample ballots × education	–0.09	0.04*	0.04#	0.08
Mailed sample ballots × age 18–24 and live with parents	0.01	0.12	0.28	0.24
Mailed sample ballots × age 18–24 and live without parents	0.33	0.13*	0.16*	0.19
Time off work for state employees	0.06	0.05	0.10	0.15
Time off work for state employees × state employee	–0.02	0.19	0.16	0.17
Time off work for private employees	–0.19	0.05**	0.07**	0.12#
Time off work for private employees × private employee	0.03	0.06	0.05	0.06
Southern state	–0.19	0.04**	0.08*	0.13
Battleground state	0.08	0.03*	0.07	0.11
Concurrent senatorial or gubernatorial contest	–0.09	0.04*	0.08	0.13

#p<0.10; *p<0.05; **p<0.01

N=44,859

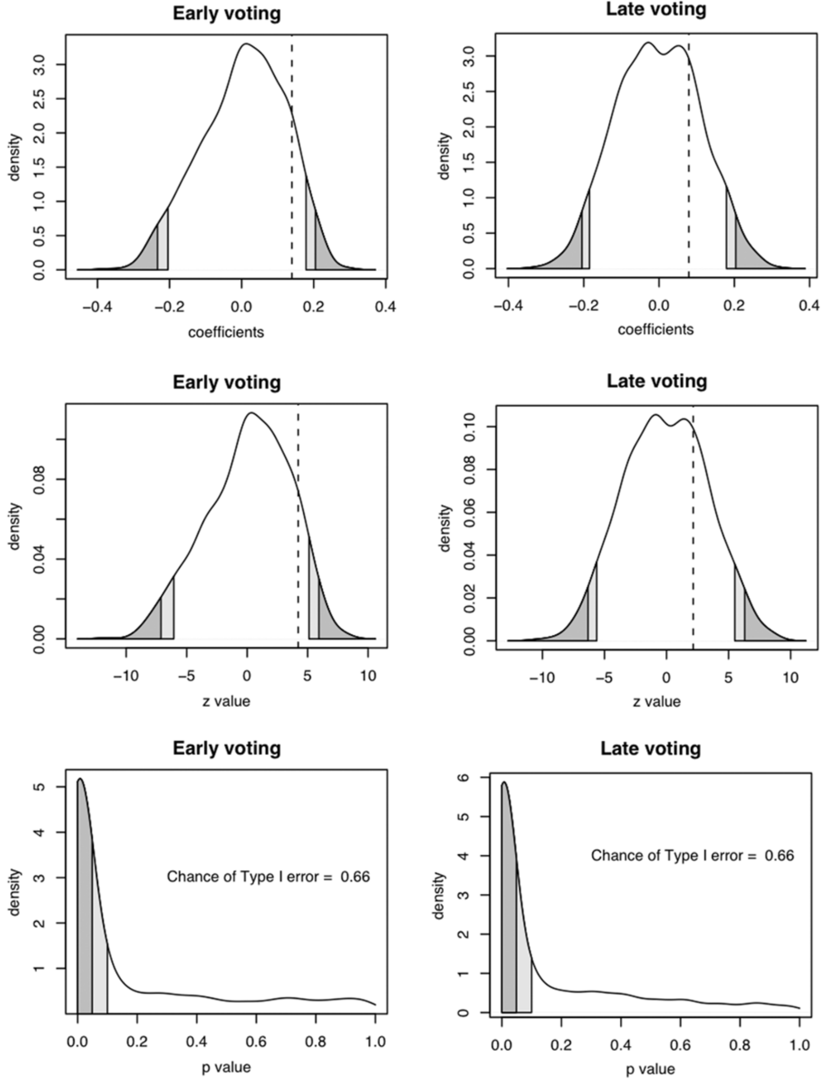
Notes: Significance levels are set at 0.10, 0.05, and 0.01, as consistent with Primo et al. (2007). Individual-level variables are omitted from the table.

Compared to the clustered standard errors, those from the randomization test are slightly larger for additive effects and considerably larger for interaction effects in most cases. The only state law effect that retains significance is the incorrectly-signed time off work for private employees (and only at the generous 0.10 level). The interaction effects between mailed polling place information and education, mailed sample ballots and education, and mailed sample ballots and young living without parents are significant using clustered standard errors but lose their significance with randomized standard errors.[5]

The best test, however, is not based on the standard errors from the randomization test, which assume a normal distribution. As discussed in the previous section, the randomization test is a distribution-free nonparametric test. For each coefficient, we can determine the portion of the random draws within 0.05 (or less) at each tail and observe whether the observed coefficient falls within these bounds.

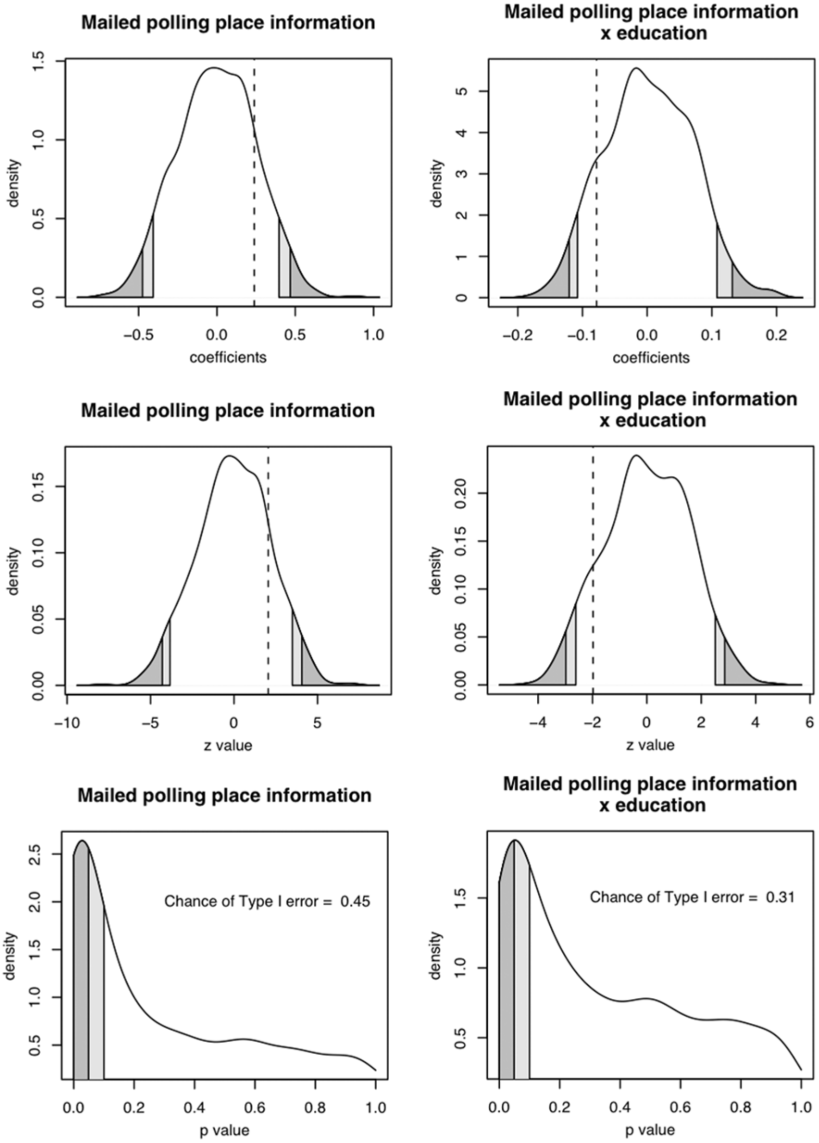
Figure 1 graphically displays some of the results from our randomization test.[6] For discussion, we take the case of the possible effect of early voting hours

Figure 1. Randomization Test Results



as our example. The three early voting graphs are shown in the first column. The first graph shows the density of the 1,000 estimated coefficients on the early voting hours indicator, estimated using 1,000 datasets in which we know no systematic relationship exists between early voting hours and turnout. As expected, the distribution is centered on zero. The dark gray shaded area represents the 5 percent most extreme coefficients, and the entire shaded area covers the 10 percent most extreme coefficients. The dotted line indicates the

Figure 1. (cont.)



Mailed polling place information

Mailed polling place information x education

Mailed polling place information

Mailed polling place information x education

Mailed polling place information

Mailed polling place information x education

magnitude of the coefficient estimated using the actual observed data, the coefficient in Table 1. From this graph, we can see that the estimated early voting hours effect is not rare by conventional statistical standards. It does not fall among the 10 percent most extreme coefficients. Therefore, we cannot rule out the possibility that the effect of early voting hours in the WHM model is just random noise. We cannot reject the null hypothesis that the early voting hours effect on turnout is zero. Again, the only state law effect that falls among the 10 percent most extreme in its randomized distribution is the incorrectly signed time off work for private employees.

The second early voting hours graph in Figure 1 shows the density of the 1,000 z-values calculated using the conventional test of significance in a multivariate logit model on the early voting hours effect. As in the coefficient graph, we see that the test statistic derived from the actual data is not among the 10 percent most extreme. For all of the state laws and interaction effects, the inferences drawn from the distribution of z values is the same as that drawn from the distribution of coefficients.[7]

The third graph in the first column of Figure 1 shows the distribution of 1,000 *p*-values associated with the 1,000 coefficients, calculated using the conventional significance test. The shaded area covers *p*-values that are 0.1 or smaller, small enough to justify rejecting the null hypothesis that early voting hours have no effect on turnout. Even though these *p*-values were calculated using data in which we know there is no systematic relationship between laws and turnout, 66 percent of the *p*-values were less than or equal to 0.1. This means that using standard significance tests with this data would cause one to falsely infer that early voting hours have an effect on turnout 66 percent of the time, instead of 10 percent of the time, as we would expect at the 90 percent confidence level. Standard tests on almost all of the post-registration laws exhibit larger than expected type I errors. This result underscores the importance of accounting for the clustered nature of the data and illustrates the challenge of testing multi-level interaction effects.

ADDITIVE MODELS

As an alternative to the complexity of the model with interaction effects involving respondents and states, we can estimate strictly additive models for subgroups that might be particularly sensitive to the stimuli of reforms designed to induce voting among the registered. Table 2 shows the results for two additive equations. The top equation is for all respondents, and it mimics Table 1 except that the interaction terms are omitted. With unadjusted standard errors, early voting and late hours appear to be highly significant. But they are not significant with clustered standard errors. Time off for private employees is the one significant reform variable with clustered standard errors and its coefficient embarrassingly has the wrong sign. When we perform our 1,000 simulations using the randomization technique, none of the reform coefficients shows up as significant.[8]

Perhaps we could find reform effects among youth, a group that might

Table 2. Additive Models for All Respondents and for Youth, 18–24 only

	Coefficient	Unadjusted Standard Errors	Clustered Standard Errors	Randomized Standard Errors
Additive Model, All Respondents				
Early voting	0.14	0.03**	0.10	0.11
Late voting	0.08	0.04*	0.08	0.11
Mailed polling place information	0.04	0.06	0.14	0.16
Mailed sample ballots	0.09	0.05	0.12	0.16
Time off work for state employees	0.06	0.05	0.10	0.15
Time off work for private employees	–0.18	0.04**	0.06**	0.13
Southern state	–0.19	0.04**	0.08*	0.12
Battleground state	0.08	0.03#	0.07	0.11
Concurrent senatorial or gubernatorial contest	–0.09	0.04#	0.08	0.13
N=44,859				
Additive model, 18–24 only				
Early voting	0.20	0.09*	0.13	0.16
Late voting	0.12	0.09	0.12	0.15
Mailed polling place information	–0.15	0.15	0.19	0.21
Mailed sample ballots	0.49	0.13**	0.20*	0.22*
Time off work for state employees	–0.16	0.12	0.15	0.21
Time off work for private employees	0.06	0.10	0.13	0.18
Southern state	–0.06	0.10	0.10	0.18
Battleground state	–0.01	0.09	0.11	0.15
Concurrent senatorial or gubernatorial contest	–0.30	0.10**	0.10**	0.18#
N=3,697				

#p<0.10; *p<0.05; **p<0.01

Notes: Significance levels are set at 0.10, 0.05, and 0.01, as consistent with Primo et al. (2007). Individual-level variables are omitted from the table.

be most receptive to efforts for improving turnout among the registered. Here, we see early voting and mailed sample ballots as significant with the unadjusted standard errors. The mailed sample ballot survives as significant both with the clustered standard errors and our randomization test with the usual 1,000 simulations.[9] Evidently youth respond to the receipt of a sample ballot.

WHY NO SIGNIFICANT EFFECTS?

How could it be that our estimates of the effects of state reforms are largely not significant? Let us use the additive model aggregated to our 42 states. Then, we go step by step to observe the gains and limitations of statistical leverage as we go from a simple aggregate (42 state model) to the contextual model.[10]

Start with the model where the units are the 42 states and the independent variables are the six reform variables. This yields Equation 1 shown in Table 3. Equation 2 incorporates controls for southern state, battleground state, and concurrent senatorial or gubernatorial contest. In each case, the results are disappointing for the reform hypothesis. Only one variable is significant in each equation, and that is the wrongly signed coefficient for time off for private employees. Extended evening hours barely reaches significance, but only without the added controls. Collectively, the six reforms have a significance level of only 0.07 (0.08 with the controls), short of the usual 0.05 benchmark.

This bare bones model shows that we should control for individual effects. Suppose we do so by adding a summary measure of the state-aggregated individual effects from the individual-level equation. For this variable, we take the prediction equation from the individual-level analysis and subtract out the estimated effects of the state-level variables. Then, we take the state means of this individual-level equation and enter them into the state-level equation. The results are shown in Equation 3.

We now have a summary control for the contribution of average individual-level effects to state turnout. Adding this variable allows us to explain half (but

Table 3. Aggregate Analyses Using U.S. States as Units

	Equation 1		Equation 2		Equation 3	
	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Early voting	0.17	0.08*	0.06	0.09	0.00	0.09
Late voting	0.04	0.13	0.13	0.09	0.10	0.09
Mailed polling place information	0.01	0.04	–0.03	0.13	0.00	0.13
Mailed sample ballots	0.19	0.12	0.03	0.04	–0.02	0.04
Time off work for state employees	–0.23	0.10*	0.15	0.12	0.22	0.12*
Time off work for private employees	0.17	0.08*	–0.27	0.10*	–0.16	0.10
Southern state			–0.22	0.10*	0.56	0.18**
Battleground state			–0.03	0.09	0.06	0.07
Concurrent senatorial or gubernatorial contest			–0.03	0.10	0.26	0.10*
Mean individual predicting from respondent characteristics					1.74	0.36***
Adjusted R^2	0.14		0.19		0.52	
Probability the six reforms are collectively significant	0.07		0.08		0.11	

N=42
*p<0.05; **p<0.01; ***p<0.001

only half) of the variance in state turnout. The important point remains that even with these controls, the reform coefficients are no more impressive than before. The control for state-level individual effects generates some churning of the estimates, but they are decidedly not very significant. Only time off for public employees passes the 0.05 threshold, but is almost offset by the wrongly-signed negative coefficient for time off for private employees. With the added controls, the net effects of the six reforms become even less significant. Controlling for state characteristics plus the individual-level characteristics of the state samples, the six reforms are significant only at the 0.11 level.

While this aggregate-level exercise is informative, it is better to estimate the effects of the specific reforms from the individual-level equation. From the individual-level equation, we can aggregate up to the state level in terms of the individual-specific characteristics, the state characteristics, and the state reforms. For this exercise, we divide the individual-level prediction equation into three components—the individual-level characteristics, the state-level traits (southern state, battleground state, senatorial or gubernatorial contest), and the important component based on the six reforms. The results are shown in Table 4.

Here, we see that all three sets of predictors are significant when aggregated to the 42 states. Of special interest is that the six predictors are collectively significant at the 0.05 level. This may be the best we can do in terms of arguing for the collective significance of the reforms on voting rates among registrants using the 2000 CPS data.

So, again, why can we not find better evidence that reforms boost turnout? One's intuition might be that once we know a state citizenry's individual characteristics and some state characteristics in terms of whether or not the state is southern, a battleground state, or has a senatorial or gubernatorial contest, that we can explain most of the factors affecting a state's turnout. It would seem, then, that with little further residual variation to control for, we can readily estimate the effects of reforms on turnout. But our surmise is incorrect. Combining mean individual characteristics (as measured), state

Table 4. Predicting U.S. State-Level Turnout from Three Components

	Coefficient	Standard Error
Individual-level predictions	1.20	0.29***
State characteristics	1.50	0.68*
State-level reforms	0.63	0.31*

Adjust R^2=0.49
N=42
*p<0.05; **p<0.01; ***p<0.001

characteristics, plus our reforms, we can do no more than explain half the aggregate-level variance of turnout among registrants in the 42 states. With turnout reforms leading to modest effects at best, it is no wonder that the reform coefficients are rarely significant.

CONCLUSIONS

We have used the nonparametric technique known as the randomization test to show that the reported effects of reforms designed to encourage turnout among registered voters are not statistically significant. Specifically, we conducted multiple simulations where the state labels were scrambled so that the distributions of the clusters of laws were assigned randomly. We applied this methodology both to the WHM interactive model and to the additive version. We also applied this to the additive version to a youth sample. Based on the simulations, the observed coefficients for state laws from the WHM analysis of turnout effects are essentially not statistically significant.[11]

The data for this study comes from a cross-section of registered voters in 2000. As a cross-sectional study, the analysis is limited by endogeneity concerns. Reforms are not distributed randomly in the states. We might expect that states with high turnout levels are most prone to pass legislation that expedites the voting process. Alternatively, it might be that state legislatures are more likely to pass legislation as a response to a sluggish voting record. These possibilities are reasons why the best design for inferring the effects of reform legislation would be some sort of a time series design.

Despite the “negative” nature of our findings, we do not argue that legislation designed to encourage turnout among registered voters is ineffectual. The various acts, such as early voting and late poll hours, might well have their intended intent. Nothing in our findings denies that reforms influence turnout by at least a few percentage points. But if they do increase turnout by only a few percentage points, the effects are difficult to estimate at the state level. Sufficient noise is present, in the form of unobserved sources of state turnout, to prevent the estimated effects from passing the usual thresholds of statistical significance.

ENDNOTES

The authors would like to thank Benjamin Highton and David Primo for sharing their data and helpful comments. They would also like to thank Robert Franzese, Eduardo Leoni, and Kevin Arceneaux, along with the three anonymous reviewers who commented

on an earlier version of this manuscript. A previous version of this article was presented at the 2008 State Politics and Policy Conference.

1. Because we focus on the modeling of state-level characteristics and their interactions with individual-level characteristics, we do not report the individual characteristics in the WHM model from our table. They include measures of employment status, education, age, income, race, and residential stability.

2. We do worry, however, that the empirical enterprise is plagued with endogeneity problems, where policy interventions to boost turnout are motivated by turnout at the state level.

3. Ideally, the permutations of the variable of interest should be sampled without replacement from the set of all possible permutations. For computational efficiency, we chose to conduct our randomization tests by sampling with replacement. Given the extremely large set of possible permutations, the chance that the same one would be drawn more than once is small. If it were to happen, the effect on the resultant reference distribution would be negligible.

4. The logic behind this recommendation is as follows. Shuffling Z does not hold constant the collinearity between Z and the other covariates X. For example, if Z and X are highly collinear, then we would expect the standard errors on the coefficients of these variables to be large. Because shuffling Z destroys the collinearity between Z and X, the coefficients obtained from the randomization method might not vary as much as they would in actual repeated sampling. Thus, inferences from the distribution of randomized coefficients would be too confident. Because test statistics are adjusted for variance magnitude, they incorporate information about collinearity in the data.

5. Throughout, we refer to the statistical significance of the coefficients on interaction effects. We are aware, of course, that the p-value associated with an interaction coefficient tests only whether the effect of one independent variable on the dependent variable depends on another independent variable. For example, the test on the interaction term between mailed polling place information and young living without parents tests whether the effect of mailed polling place information on a registered voter's propensity to vote is different if the person is young and living without parents versus if he or she is not. This test remains only one of many interactive hypotheses one could conduct using a model like that in WHM (see Kam and Franzese 2007). Furthermore, one cannot infer the direction of this conditioning effect by observing the sign of the interaction coefficient because this is a logit model (Ai and Norton 2003). However, because proper calculation of the standard errors is our primary concern, we do not pursue the investigation into the marginal effects of laws any further. Suffice it to say, the randomization standard errors are, as expected, larger than the clustered standard errors, and this expansion makes a difference for at least one type of interactive hypothesis.

6. Full graphical results are available upon request.

7. This is probably because we shuffle the menu of state laws, instead of each law individually, which preserves the collinearity among the laws.

8. The five coefficients (excluding time off for private employees) are collectively significant, however, at the 0.03 level.

9. Using clustered standard errors, the collective significance of the six reforms is a weak 0.34. Results this far from zero could have occurred one time in three if all six null hypotheses are true.

10. Given the lack of scholarly consensus on how many groups are needed for proper inference in multi-level data, Angrist and Pischke (2009) recommend aggregation as a robustness check on group-level effects found using individual-level data.

11. The WHM model uses data on individual voters, with state-level variables as contextual variables. Two further alternatives are to employ multilevel modeling (with random effects) and to employ fixed effects where state dummy variable coefficients from the individual analysis are modeled in terms of state characteristics. We anticipate that randomization tests for these models would also find reform effects outside the usual bounds of significance.

REFERENCES

Ai, Chunrong, and Edward C. Norton. 2003. "Interaction Terms in Logit and Probit Models." *Economics Letters* 80:123–9.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arceneaux, Kevin, and David W. Nickerson. 2009. "Modeling Certainty with Clustered Data: A Comparison of Methods." *Political Analysis* 17:177–90.

Donald, Stephen G., and Kevin Lang. 2007. "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics* 89:221–33.

Donohue, John J., and Justin Wolfers. 2006. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review* 58:791–835.

Edgington, Eugene S., and Patrick Onghena. 2007. *Randomization Tests*. 4th ed. Boca Raton, FL: Taylor and Francis Group.

Fisher, R.A. 1935. *The Design of Experiments*. Edinburgh, UK: Oliver and Boyd.

Griffin, John D., and Michael Keane. 2006. "Descriptive Representation and the Composition of African-American Turnout." *American Journal of Political Science* 50:998–1012.

Helland, Eric, and Alexander Tabarrok. 2004. "Using Placebo Laws to Test 'More Guns, Less Crime.'" *Advances in Economic Analysis and Policy* 4:1–7.

Hogan, Robert E. 2005. "Gubernatorial Coattail Effects in State Legislative Elections." *Political Research Quarterly* 58:587–97.

Kam, Cindy J., and Robert J. Franzese, Jr. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis.* Ann Arbor, MI: University of Michigan Press.

Keele, Luke, Corrine McConnaughy, and Ismail White. 2008. "Statistical Inference for Experiments." Unpublished paper. Available at www.polisci.ohio-state.edu/faculty/lkeele/randtests.pdf (January 4, 2010).

Kennedy, Peter E. 1995. "Randomization Tests in Econometrics." *Journal of Business and Economic Statistics* 13:85–94.

Kennedy, Peter E., and Brian S. Cade. 1996. "Randomization Tests for Multiple Regression." *Communications in Statistics—Simulation and Computation* 25:923–9.

Lawless, Jennifer L. 2004. "Politics of Preference? Congresswomen and Symbolic Representation." *Political Research Quarterly* 57:81–99.

Leoni, Eduardo L. 2009. "Analyzing Multiple Surveys: Results from Monte Carlo Experiments." Unpublished paper. Available at http://eduardoleoni.com/workingpapers/multilevel.pdf (January 4, 2010).

Manly, Bryan F. J. 1997. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. 2nd ed. London, UK: Chapman Hall.

Moore, David S., George P. McCabe, William M. Duckworth, and Stanley L. Sclove. 2003. *The Practice of Business Statistics Companion Chapter 18: Bootstrap Methods and Permutation Tests*. New York, NY: W.H. Freeman.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32:385–97.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables in Micro Units." *Review of Economics and Statistics* 72:334–8.

O'Gorman, Thomas W. 2005. "The Performance of Randomization Tests that Use Permutations of Independent Variables." *Communications in Statistics—Simulation and Computation* 34:895–908.

Primo, David M., Matthew I. Jacobsmeier, and Jeffrey Milyo. 2007. "Estimating the Impact of State Policies and Institutions with Mixed-Level Data." *State Politics and Policy Quarterly* 7:446–59.

Soss, Joe, Laura Langbein, and Alan R. Metelko. 2003. "Why Do White Americans Support the Death Penalty?" *The Journal of Politics* 65:397–421.

Wolfinger, Raymond E., Benjamin Highton, and Megan Mullin. 2005. "How Post-registration Laws Affect the Turnout of Citizens Registered to Vote." *State Politics and Policy Quarterly* 5:1–23.

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who Votes?* New Haven, CT: Yale University Press.