CALTECH/MIT
VOTING TECHNOLOGY PROJECT

A multi-disciplinary, collaborative project of
the California Institute of Technology – Pasadena, California 91125 and
the Massachusetts Institute of Technology – Cambridge, Massachusetts 02139

STUDYING ELECTIONS: DATA QUALITY AND PITFALLS IN MEASURING THE EFFECTS OF VOTING TECHNOLOGIES

R. Michael Alvarez
Caltech

Stephen Ansolabehere
MIT

Charles Stewart III
MIT

Key words: ***data quality, data availability, residual votes, turnout, statistical consequences***

VTP WORKING PAPER #21
November 2004

Studying Elections:

Data Quality and Pitfalls in Measuring the Effects of Voting Technologies

Introduction

Professor Geralyn Miller reminds us of the range of voting administration practices across the United States. We use this variability to study the average performance of various types of voting equipment throughout the country (Ansolabehere and Stewart, forthcoming). Professor Miller suggests that the performance of equipment is, in fact, quite variable across states. A particular technology that performs poorly nationwide might perform well in a particular setting --- either because the technology is well suited to the peculiarities of the setting or because a locality has been proficient in overcoming shortcomings that vex other jurisdictions. In making this point, Professor Miller examines two states, Wyoming and Pennsylvania, in the 2000 election.

While we are sensitive to the general point Miller makes, her paper does not in fact demonstrate it. Instead, careful consideration of this paper raises a separate, but equally important matter – the content and quality of local and state election reports. The data she employs run up against problems that face all researchers doing this type of analysis. Rather than mount a full-scale critique of Miller's findings, we think it more constructive to focus on the two major data problems in her paper, as an illustration of precisely how hard it is to conduct this type of research. The most serious errors in Miller's paper are not readily apparent to most

researchers doing voting technology research. Indeed, as we will show, Miller stumbled upon one error that we committed and were publicly chastised for.

The two states that Miller studies illustrate separate and important data problems. Pennsylvania illustrates that states do not report all the data necessary to study the performance of voting technologies. Wyoming illustrates that not all states report all they seem to be reporting.

Beyond data collection concerns, there is also a basic issue of research design. Single cross-sectional studies of individual states in fact have little statistical power. The number of counties is simply too small to have arrive at meaningful estimates of average effects of technologies, let alone the interactive or varying effects of technologies used. Research on election administration needs to go beyond looking at single elections in order to establish the point that voting technology performance varies across states. That lesson is most clearly borne out in our prior research on this subject, in which many puzzling results emerge in cross-sections that are resolved in panel studies.

Methodology Matters

Important methodological issues arise in the study of voting technologies. The first is that one must understand exactly how the data were generated. We leave to the next sections consideration of the data reported by the states and counties in each of the two cases Professor Miller examines.

Second, study of county-level data within a state quickly run up against the boundary of statistical power. All estimates of machine type differences suggest that the difference in performance between the "best" and "worst" voting machines may be in the order of 1% of all ballots cast. While this may seem like a small percentage, it represents a significant fraction of

the quadrennial variability in national turnout. Any single factor that would increase or decrease turnout by one million voters nationwide, which is approximately 1% of turnout, would be noticed.

In a single state with a small number of counties, a 1% difference is usually too small to be discerned with any degree of statistical power. At the tradition level of $p < .05$, the standard error must be .005 for a 1% difference to be judged statistically significant in a *t*-test (to take the simplest example). Yet, for such a small standard error to be achieved, the number of observations must be substantial. Consider a simple example. Technology A has a residual vote rate of 1.5% while Technology B has a residual vote rate of 2.5%. In both sets of counties, the standard deviation of the residual vote rate is 2%. Assume the simplest case, where we have an equal number of observations on counties with Technology A and Technology B. How large does the sample need to be, in order to produce a standard error of .005? The answer is approximately 1,600. That is significantly larger than the number of counties in Wyoming (23) or Pennsylvania (67). In the case of Wyoming, the differences between the different voting technologies in a single year would have to be scandalous beyond belief to survive a statistical test with such a small number of observations.

To overcome the small-*n* problem, researchers have resorted to a number of strategies. The first is simply to increase the number of observations. Miller's use of precinct data is clearly an attempt to do this. Unfortunately, as we note later, the precinct data she used did not include the write-in vote for Nader. This leads to bias in the estimates; increasing sample size will not resolve this bias. Regression analyses across precincts offers some help to the extent that the control variables correct the particular source of bias. Regression models can also improve precision, since the control variables capture variation in the dependent variable. Unfortunately,

the demographic or political controls that might be included in a regression are rarely if ever reported at the precinct level.

A better strategy for getting at state (or national) differences requires less data than precinct data --- create a panel. In other words, the spirit of Miller's article can be maintained by including residual vote rates from all Wyoming counties across a number of years. If the time period of the panel is short enough to allow us to assume that the county-level independent variables that influence residual votes (other than voting machines used in year *t*) are constant, then we can employ *fixed effects regression.* Under fixed effects regression, the omitted independent variables are accounted for with a series of dummy variables, one for each county (in this case). In addition, because there is year-to-year variability in aggregate residual vote rates, we include a dummy variable for each election year. Finally, the test for voting machine effects is conducted with a series of dummy variables, equal to 1 if county *c* used a particular machine in election year *t*, 0 otherwise.

Table 1 reports the analysis of a fixed effects regression of Wyoming, using the data we previously gathered, or purchased from Election Data Services, Inc.[1] Like Miller, we excluded Big Horn County, because it used a mix of equipment. We also weighted each observation by turnout, so that the results could be interpreted in terms of the probability that an individual voter would cast a residual vote.

[Table 1 about here]

The three dummy variables for years simply allow residual vote rates to vary by year, most likely because the percentage of intentional abstentions fluctuates because of differences in

---

[1] Election Data Services, Inc. (EDS) is a primary provider of data on election administration; see http://www.electiondataservices.com/home.htm.

candidates and campaign intensity. Here, we see that the residual vote rates in 1992, 1996, and 2000 were between 2.2% and 3.9% lower than they were in 1988.

The dummy variables for machine type use punch cards as the omitted category, since it was the most commonly used voting machine in Wyoming during this period. Keep in mind that most counties changed their voting technology type at some point during this period: only seven of the 22 counties included in this analysis used in the same voting technology in 2000 that they used in 1988. Therefore, the dummy variables for voting technology type are leveraging off two types of information: differences in residual vote rates *across machine types* in any given year and differences in residual vote rates *across time* in counties that changed machine type during this period.

The voting technology dummy variable coefficients are all negative, indicating that lever machines, optical scanners, and DREs all averaged lower residual vote rates than punch cards. The optical scan and DRE differences are statistically significant, and at 1.9%, are substantively important. In 2000, 28,276 Wyomingites cast their ballots using punch cards. If those voters had been allowed to vote on optical scanners or DREs, these results suggest that 537 more votes would have been recorded for president in Wyoming, simply by virtue of upgrades to voting technology.

In the analysis we have conducted in other papers, we have used fixed effects regression to estimate differences across voting machine types nationwide (Ansolabehere and Stewart, forthcoming). In these regressions, based on over ten thousand observations, we are able to include other controls, such as state dummies, to account for differences in administrative practices across states, and turnout, to account for differences when large numbers of voters unexpectedly surge into the electorate. Because of the large number of observations, the results

we get in those analyses are much more precise. However, they are substantively different from what we get when we analyze Wyoming alone. For instance, in Ansolabehere and Stewart (forthcoming, Table 3), we find optical scanners to have residual vote rates 1.3% less than punch cards and DREs to have residual vote rates 0.6% less than punch cards. The Wyoming differences appear to be larger in magnitude and slightly different in substance. However, all these analyses share one thread: punch cards are the worst, nationwide and in Wyoming.

Researchers must be very cautious about basic questions of statistical power when they study elections and voting technology, or indeed when they use any county-level data within a single state. An equally fundamental issue, though, concerns the quality of the data employed in a given study. Understanding of the problem with Professor Miller's analysis emerges upon careful consideration of the two states under study.

Pennsylvania: States don't always report what we need

In our research, we examined voting technology performance nationwide from 1988 to 2000 at the county-level (Caltech/MIT Voting Technology Project, July 2001). We immediately discovered that many states do not require their counties to report all the data necessary to conduct basic research into voting technology performance. As we document in our recent recommendations to the Election Assistance Commission (EAC), there are at least eleven states that fit into that category (Caltech/MIT Voting Technology Project October 2004). This is a major problem for auditing and inspecting elections in these states, as even the most rough and ready measures of performance, residual votes, cannot be computed.

Without a statewide certification of turnout, many problems and inconsistencies in the data reported in a state might arise. In states where no statewide certification of turnout is made, counties have considerable discretion in deciding what to report for total ballots cast. Sometimes

counties make a complete report of total ballots cast. However, we also discovered in our data collection efforts that, if there is no official certification, counties sometimes report the election night tally (which may not even include all precincts), the total ballots cast in precincts (which does not include absentees), or nothing at all. [2]

Pennsylvania is one of the eleven states that have no statewide turnout certification. It is also one of two cases Professor Miller has chosen to study. In our data collection efforts in 2001, we were able to track down information on turnout from five Pennsylvania counties (of 67) --- Beaver, Cambria, Chester, Lehigh, and Washington. Some of these counties were of particular interest because of the technologies used. However, because of questions about the certification process, we left this state out of the overall data analysis. Pennsylvania itself consisted of only a small fraction of the sample. Our entire panel data set amounted to 13,000 observations, consisting of county-level election returns from 1988 to 2000. The data do include states similar to Pennsylvania in many respects, such as New York, Ohio, and Indiana. For the purposes of estimating *nationwide* voting technology performance, the Pennsylvania data were not necessary; nor did we have much confidence in what we were able to collect.

Miller made a more concerted effort gathering data from Pennsylvania. She contacted each county in Pennsylvania and obtained the number of total ballots cast that each reported. She deleted three counties that could not produce total ballots cast, and six counties that reported

---

[2] Following the 2002 election, we employed a team of research assistants to gather election and machine data, to follow up on our previous research. A significant fraction of the team's time was spent contacting local officials in Pennsylvania and other states that did not certify total ballots cast. Many of these offices could not provide information requested and were frequently hostile toward the students and faculty making the request.

using a mixture of voting systems for precinct voting. She provided us with her data, which we use here (and below) to explore some implications of her arguments.

We begin with a simple attempt at replication.

The first step in the replication was to look at the data. The most striking thing about the data set is that three counties --- Dauphin, Mercer, and Philadelphia --- had *negative* residual vote rates in the presidential race. That is, they had more votes recorded for president than they had voters recorded as voting.

We checked the presidential vote data in Miller's dataset against that provided online by the Pennsylvania Department of State. Unfortunately, this did not clarify the matter much.[3] In two cases (Dauhpin and Mercer Counties), the Department of State website actually reported an even greater number of ballots counted in the presidential race, thus making the residual vote even more negative. In the case of Philadelphia County, Miller reports 563,339 presidential votes counted, while the Department of State website reports 561,180 --- but both of these numbers are still greater than the number of reported total ballots cast used by Miller (560,179).

Why do such discrepancies occur? In our experience, they arise from many different sources – failure to include absentee ballots, use of preliminary vote tallies to get information out to the media, and typographical errors in state or county reports. In states that do not have statewide certification of total ballots cast, there are often large discrepancies across counties in what is reported, as each county uses a different standard to establish turnout.

In one respect these sorts of errors may not be catastrophic. If the underlying residual vote rate is highly correlated with the observed rate calculated, regressions and differences of

---

[3] http://web.dos.state.pa.us/cgi-bin/ElectionResults/county2.cgi?eyear=2000&etype=G&office=USP accessed on Sept. 9, 2004.

means may still capture the effects of technology. Nonetheless, in our research, we have dropped counties with negative residual vote rates.

The second step in the replication was to reconstruct Table I in Miller's paper, which reports the correlation coefficients between voting technology type and residual vote for president and senator (measured as a percentage of the vote). We were unable to do so. (See Table 2.) The reason is simple: Miller uses the wrong total turnout figure for Bucks County. In an original data set provided to us by Professor Miller, the turnout figure for Bucks County was erroneously reported to be 564,471. When our problems replicating Table 1 were brought to Professor Miller's attention, she informed us that the correct turnout figure for Buck County was 264,471. However, she apparently never re-calculated the presidential and senatorial residual vote rates using the corrected Bucks County figure. In other words, Table I in the Miller paper can only be replicated if we use the original, incorrect, number for turnout in Bucks County. As far as we can tell, the results in Table I are due to a typographical error.

[Table 2 about here]

Using the corrected turnout figure for Buck County turnout, we were similarly unable to replicate Table II, which reported regression results predicting residual vote rates for senator and president.

Wyoming: Errors and Omissions

Miller's second case is Wyoming. Wyoming state law requires counties to report turnout, in addition to vote counts. On the face of it, Wyoming appears to be an exemplary case of how state officials can report the data necessary for their citizens, and others, to judge the performance of their election technologies. Immediately after the 2000 election, the Wyoming Secretary of State posted precise information about the voting systems used by each county, the

total number of ballots cast in each county, and the votes received by each candidate on the ballot.[4] Based on the available information, our July 2001 report listed Wyoming with a 3.6% residual vote in the 2000 presidential election, the second highest among the states we had data for, behind Illinois.

Shortly after our July 2001 report was published, communications between the Wyoming Secretary of State and our university presidents ensued, with the Wyoming officials claiming that our analysis of their state's residual vote rates was flawed. In the exchange that ensued, the source of the problem became evident.

Ralph Nader was not on the Wyoming ballot, but he received a significant write-in vote. Under Wyoming law, the votes of write-in candidates are included in the statewide canvass *only* if they affect the outcome of the election.[5] Therefore, Nader's vote appears to be part of the residual vote in the Wyoming web site. However, because Nader had formally requested that Wyoming record his write-in votes, the Elections Division had gathered this information from the county election directors. It was available for anyone who asked, but you had to know to ask. Because we did not know to ask, we didn't. We assumed that the certification of the vote was of all votes cast.

---

[4] It appears that Wyoming added even more detailed information about election returns to their web site, after our data gathering sweep in early 2001. Of particular note is the precinct-level data that Miller uses in her paper. According to the technical information about the web content on that site, the precinct data were added on February 5, 2001. Our initial gathering of county-level information occurred on January 18, 2001.

[5] Wyoming Statute, 22-16-103; personal correspondence with Peggy Nighswonger, Wyoming Election Director, 7 August 2001.

After re-examination, we issued an errata for our 2001 report, in which we noted that "The residual vote rate listed for the state of Wyoming for the year 2000 should be 1.5," acknowledging that the write-in votes in the 2000 presidential race were substantial enough to influence our residual vote estimates.[6]

This omission (which appears to be unique among the states) is critical for judging the relative performance of voting technologies in Wyoming.[7] Variability in the Nader vote in Wyoming in 2000 is correlated with the type of voting technology used. Counties that had a high write-in vote for Nader were significantly more likely to use optical scanners than those with few Nader votes. (See Table 3.) Consequently, this had the effect of making it seem like optical scanners performed less well in Wyoming in 2000 than in the rest of the nation, and that optical scanning performed less well in Wyoming than other technologies. In fact, when we correct for the Nader vote in the various counties, the relative performance of optical scanners (and lever machines), compared to other equipment, is also consistent with the nationwide trends we have reported.

[Table 3 about here]

To be clear about this: Miller utilizes precinct-level data from Wyoming that is *precisely* the data we originally used, which omitted the Nader vote. Precinct level data can be very valuable in this sort of analysis; in this case, it is valuable *only* if it includes the Nader write-in votes in each precinct. It does not. Therefore, Miller's conclusions are not valid.

---

[6] See http://vote.caltech.edu/Reports/2001report.html

[7] Wyoming announced in September 2004 that it would publicly report write-in votes for president in 2004. See Thomas Hargrove, "Wyoming Agrees to Change Election Reporting," Scripps Howard News Service, September 16, 2004, accessed via LexisNexis.

In fact, if Miller had the precinct data corrected for the Nader vote, it would follow almost precisely the results in Table 3. That is because the right way to measure differences in residual vote rates across voting technologies is to weight each observation (whether precinct or county) by turnout. In cases where all the precincts in a county use the same technology, it does not matter whether we aggregate up from the precinct or county level --- the weighted averages will be identical. In the case of Wyoming, there is one county that used two methods of voting; knowing which precinct used which method allows for some differences in the two different aggregation methods. In the case of Wyoming, however, the difference is trivial.

Conclusions

Understanding the performance of voting technologies in the United States using scientific methods is a critical part of the policy process that tries to improve voting in America. Previous research we have conducted has helped shape national and state policy in technology upgrades. Professor Miller suggests that each state may in fact deviate from the national average, and that some technologies might be most appropriate for particular states. Punch cards might work well in Pennsylvania; optical scanning might work poorly in Wyoming.

We have analyzed the data presented in Professor Miller's article, and conclude that the data in fact support our findings concerning the weaknesses of punch cards overall and the comparatively good performance of scanners. In Wyoming, using the correct accounting of presidential ballots cast, optical scanning performs much better than punch cards or electronics. The figures are remarkably close to the national pattern. In Pennsylvania, because the state does not provide for a certification of the total ballots cast, the appropriate data for the assessment of voting technology do not appear to be available, even after Professor Miller's considerable

efforts to collect the appropriate data. Some counties show negative residual votes, which are logically impossible.

The more general lesson we draw from our research, though, is that there are significant data challenges in the field of election administration and election studies. As we have noted in two recent reports (Caltech/MIT Voting Technology Project July 2004, October 2004) there is a disturbing lack of detailed election administration data available for studying elections. In past elections, some states have not reported the number of total ballots cast at the county level, others have not fully reported votes cast for all candidates (including write-in votes) in federal races, and in many situations we do not know what types of voting technologies have been used by voters. The situation is even more problematic when it comes to other aspects of the electoral process, as we rarely can find data on very important aspects of the process like the numbers of absentee and provisional ballots cast, and the numbers of absentee and provisional ballots rejected.

More needs to be done to help state and local election officials understand the importance of collecting and making election data available to the public after each federal election. In particular, election officials should collect and distribute (at the lowest level of aggregation possible):

1. The total number of registered voters who cast ballots in the jurisdiction
2. The total number of votes cast for every federal candidate on the ballot, including all votes cast for write-in candidates
3. The numbers of absentee, early, and provisional ballots distributed, and the numbers of these included in official vote tabulations
4. The technologies utilized for precinct, early and absentee voting

5. Documentation about all incidents and problems that arose regarding voting equipment and administrative issues, including the steps taken to resolve these problems.

Provision of these data (and other more detailed data when possible) would help the public and policymakers understand the elections process in better detail and help to insure the integrity of future elections.

References

Ansolabehere, Stephen and Charles Stewart, forthcoming. "Voting Technology and Lost Votes." *Journal of Politics.*

Caltech/MIT Voting Technology Project, 2001. "Voting: What Is, What Could Be." http://vote.caltech.edu.

Caltech/MIT Voting Technology Project, 2004. "Immediate Steps to Avoid Lost Votes in the 2004 Presidential Election: Recommendations for the Election Assistance Commisssion." http://vote.caltech.edu.

Caltech/MIT Voting Technology Project, 2004. "Insuring the Integrity of the Electoral Process: Recommendations for Consistent and Complete Reporting of Election Data." http://vote.caltech.edu.

Table 1. Fixed effects regression predicting residual vote in Wyoming presidential elections, 1988-2000.

Year effects (1988 comparison category):	
Year=1992	-0.039
	(0.004)
Year=1996	-0.022
	(0.006)
Year=2000	-0.025
	(0.007)
Voting technology effects (punch cards comparison category)	
Lever machines	-0.010
	(0.011)
Optical scanners	-0.019
	(0.007)
DRE	-0.019
	(0.009)
Constant	0.057
	(0.004)
N	88
R^2	.72
F-test for 22 county dummy variables (coefficients suppressed) (d.f.=22,63)	1.03

Table 2. Correlation coefficients for Pennsylvania Equipment and Residual Votes: Replication of Miller Table I.

Office		Paper	Lever Machines	Punch Cards	Optical Scan	DRE
U.S. Senate	Miller	-.038	.383	-.120	-.242	-.047
	Replication	-.035	.278	-.089	-.167	-.042
U.S. President	Miller	-.043	.199	-.051	-.159	.040
	Replication	-.037	.193	-.057	-.129	-.000

Table 3. Residual vote rate and Nader vote, by type of voting technology used in Wyoming, 2000 presidential election.

	Nader pct.	Residual vote rate w/out Nader vote	Residual vote rate with Nader vote
DRE	0.8	2.8	2.0
Lever machine	0.6	1.9	1.3
Mixed	0.9	3.3	2.4
Punch card	0.7	3.2	2.5
Optical scan	2.4	3.8	1.3
Total	2.1	3.6	1.5