COUNTING ALL THE VOTES:

The Performance of Voting Technology in the United States

Henry E. Brady, Justin Buchler, Matt Jarvis, John McNulty
Department of Political Science, Survey Research Center, and
Institute of Governmental Studies, University of California, Berkeley

September 2001

TABLE OF CONTENTS

EXECUTIVE SUMMARY 1

STUDY GOAL
DATA QUALITY
USING RESIDUAL VOTES AS THE MEASURE OF PERFORMANCE
METHODOLOGICAL ISSUES
PRESIDENTIAL RESIDUAL VOTES AND VOTING SYSTEMS
SENATORIAL AND GUBERNATORIAL RESIDUAL VOTES AND VOTING SYSTEMS
RECOMMENDATIONS

INTRODUCTION 4

I. DEFINING PERFORMANCE 8

MINIMIZING RESIDUAL VOTES
OTHER WAYS TO EVALUATE VOTING SYSTEMS

II. TYPES OF VOTING SYSTEMS AND THEIR SUSCEPTIBILITY TO MISTAKES 10

FIVE TYPES OF SYSTEMS
Paper Ballots
Mechanical Voting Machines ("Lever Systems")
Punchcard Systems
Mark-sense or Optical Scan Systems
Direct Record Electronic Systems
Absentee Ballots
SYSTEM FEATURES AND THE SOURCES OF RESIDUAL VOTES

III. DATA COLLECTION METHODOLOGY 17

DATA REQUIREMENTS
VOTING SYSTEM DATA
ELECTION DATA
DEMOGRAPHIC AND POLITICAL DATA
THE FINAL DATA SET
COMPARISON WITH THE CAL-TECH/MIT DATA

IV. STATISTICAL ANALYSIS—THEORETICAL ISSUES 22

THE PITFALLS OF COMPARING AVERAGE RESIDUAL VOTES FOR VOTING SYSTEMS
SYSTEM PERFORMANCE MAY VARY BY COUNTY CHARACTERISTICS
HOW MANY UNITS?
CONTROLLING FOR THE DIFFERENT CHARACTERISTICS OF COUNTIES

V. DESCRIPTIVE STATISTICAL ANALYSIS FOR PRESIDENTIAL RESIDUAL VOTES 29

AVERAGE PERFORMANCE IN ALL COUNTIES
POORLY PERFORMING SYSTEMS
GOOD, ADEQUATE, WORRYING, AND UNACCEPTABLE PERFORMANCE IN ALL AND IN LARGE COUNTIES

VI. MULTIVARIATE ANALYSIS FOR PRESIDENTIAL RESIDUAL VOTES 33

Correcting for Confounding Variables
Presidential Residual Votes in All and in Large Counties

VII. SENATORIAL RESIDUAL VOTES 36

VIII. PRESIDENTIAL RESIDUAL VOTES AND VOTER EDUCATION LEVELS 37

IX. FINDINGS ABOUT SPECIFIC SYSTEMS 42

X. TWO CALIFORNIA CASE STUDIES 44

XI. COUNTY ACQUISITION DECISIONS AND SELECTION BIAS 46

SUMMARY OF FINDINGS AND RECOMMENDATIONS 47

Goal and Purpose of the Study
Data Quality
Using Residual Votes as the Measure of Performance
Methodological Issues
Presidential Residual Votes and Voting Systems
Senatorial and Gubernatorial Residual Votes and Voting Systems
Caveats
Recommendations

APPENDIX 1: ADDITIONAL PECULIARITIES IN THE DATA SET 51

REFERENCES 52

COUNTING ALL THE VOTES:

The Performance of Voting Technology in the United States

Henry E. Brady, Justin Buchler, Matt Jarvis, John McNulty
Department of Political Science, Survey Research Center,
and Institute of Governmental Studies, University of California, Berkeley

EXECUTIVE SUMMARY

Study Goal

In response to national concerns about voting systems, the Survey Research Center and the Institute of Governmental Studies at the University of California, Berkeley have worked together[1] to determine which voting systems do the best job of recording and tabulating votes by minimizing the residual vote percentage—the percentage of ballots in a jurisdiction for which no presidential votes are recorded. Residual votes include undervotes where no vote is counted for a race and overvotes where more than one vote invalidates the ballot for that race. We compare five different voting systems: direct record electronic (DRE), lever machines, optical scan, paper ballots, and punchcards.

Data Quality

Our analysis is based upon data for 2,219 U.S. counties for the 2000 general election. This information includes the type of voting system, residual votes computed from turnout and votes for candidates, and demographic and political characteristics of the counties. We invested heavily in collecting and cleaning these data, and we believe that they provide a solid basis for our analysis. Nevertheless, we still have some concerns about their quality. These concerns with the 2000 data led us to reject the idea of using data for years before 2000 because they could not be properly audited with the same thoroughness as the 2000 data.

Using Residual Votes as the Measure of Performance

Our report focuses on the degree to which voting systems record votes for every voter who comes to the polls. Although some voters who turn out decide intentionally not to vote in the presidential election, a substantial majority of presidential residual votes are unintentional (i.e., mistaken) undervotes and overvotes. Consequently, a residual vote measure for presidential elections can provide information about the performance of voting systems. Residual votes for senatorial and gubernatorial elections are much less useful measures of voting system performance because so many of them may be intentional undervotes whose numbers differ substantially from one jurisdiction to another depending upon the closeness of the election.

Methodological Issues

Although we present significant findings in our report, we have erred on the side of modest claims because of the difficulties of drawing conclusions from non-experimental,

[1] Funding for graduate students and data acquisition was provided by Sequoia Voting Systems of Oakland, California who sell punchcard, optical scan, and Direct Record Electronic (DRE) voting systems. The Survey Research Center contributed funding to support the principal investigator. The conclusions are entirely those of the SRC/IGS research team that undertook the study.

observational, data. These difficulties are exacerbated by four features of our study. The national data are limited in significant ways described in the report. There are great challenges in adequately controlling for factors other than voter system type that might affect performance. Selection effects can further confound the results because older voting systems can gain an advantage by remaining in situations where they perform well while being replaced in situations where they work badly. Newer systems, however, have not yet found the niche where they perform best. Finally, characterizing diverse and sometimes rapidly changing technologies pose additional problems.

Presidential Residual Votes and Voting Systems

We find that DRE, lever machines, optical scan, and paper ballots *all* produce significantly fewer residual votes, between 1/2% to 1% less on average, than punchcards. This result remains true when the data are analyzed with a variety of statistical controls.

Using the four point scale devised by the National Commission on Federal Election Reform which uses the gradations of "good," "adequate," "worrying," and "unacceptable" performance, we find that all voting systems have a substantial number of unacceptable and worrying implementations, but punchcard systems are typically in the bottom two categories while all other voting systems—electronic, lever machines, optical scan and paper ballots—are typically in the top two categories.

We find, for example, that in the 100 largest counties in the United States comprising about 40 percent of the U.S. population, optical scan and electronic systems can be given the highest rating of "good" on the four point scale devised by the National Commission on Federal Election Reform (the Carter/Ford Commission). Lever machines are "adequate" by this definition and punchcards are "worrying". Paper ballots are not used in large counties. We also find that DREs and paper ballots are substantially less prone to what the Cal-Tech/MIT Voting Technology Project calls "very high residual vote rates" (5% residual vote or higher) than any other system.

Optical scan and DREs appear to dominate all other systems in terms of overall performance across all counties and especially in large counties. DREs do not do as well as optical scan in smaller counties, but optical scan systems are more prone to very poor performance than DREs. Lever systems are competitive with optical scan systems and DREs across all counties but they seem prone to very poor performance and they do not do well in large counties. Paper ballots do well, but they are only used in small counties. Punchcard systems always perform the worst.

The available evidence does not provide any substantiation for Cal-Tech/MIT's claim that DREs might be harder to use than other systems. In fact, we present evidence suggesting that DREs pose fewer problems for poorly educated voters than all other systems except paper ballots. Punchcards appear to be very error prone for everyone. Even voters with more education have trouble with them.

Based upon our own analysis of data for the 2000 election and our concerns with these problems, we advise against the rush to judgment of the Cal-Tech/MIT Voting Technology Project with respect to the relative performance of non-punchcard systems, and we agree with the Commission on Federal Electoral Reform and the National Task Force on Election Reform that much more testing of the available systems is needed before making any final judgments about the suitability of one system over another.

Senatorial and Gubernatorial Residual Votes and Voting Systems

We are very skeptical about using senatorial or gubernatorial residual votes to evaluate voting systems because of the large number of intentional undervotes. Our analysis of senatorial residual votes shows that lever and punchcard systems perform poorly and other systems perform better, but our analysis of senatorial residual votes is very sensitive to the specification of the model suggesting that our skepticism about the meaning of senatorial residual votes is justified.

Recommendations

We recommend the following with regard to voting systems in the United States:

1. Concerted efforts should be made to move away from Votomatic-style punchcard systems to other, more appropriate ones. In most cases this will mean either optical scan or electronic systems.

2. States should implement uniform reporting standards for county reporting of undervotes and overvotes. The Office of Election Administration of the Federal Election Commission should take the lead in developing these standards. At the very least, information about undervotes and overvotes should be reported by precincts and by absentee, early voting, and election day voting after every election. States should also keep up to date records on voting systems in each county or township.

3. More effort should be put into human factors engineering to find out what features of voting systems cause unintentional undervotes and overvotes.

4. Voting systems should be studied using experimental methods which vary their features to see which ones have the most impact on their performance.

In addition, we recommend the following regarding research on voting systems:

1. More and better data should be collected at the county and precinct level on:

   - The types of voting systems and their implementation,
   - The process of voter education, and
   - The level of staffing and resources devoted to election administration.

2. A national program of experimentation with voting systems should be undertaken that involve experts in engineering, human factors, psychology, political science, and economics who would use the best available experimental methods.

COUNTING ALL THE VOTES:

The Performance of Voting Technology in the United States

This is the final report on voting systems done jointly by the Survey Research Center and the Institute of Governmental Studies at the University of California, Berkeley under the direction of Professor Henry E. Brady. The study was partly funded by Sequoia Voting Systems of Oakland, California who sell punchcard, optical scan, and Direct Record Electronic (DRE) voting systems,[2] but the conclusions are entirely those of the SRC/IGS research team that undertook the study.[3] The study was begun in late May, 2001 and finished in early September, 2001.

The goal of this study is to determine which voting systems do the best job of recording and tabulating votes. Our primary measure of performance is the percentage of ballots in a jurisdiction for which no presidential votes were recorded. This "residual vote" percentage is comprised of unmarked ballots plus overvoted ballots plus uncounted ballots.[4] We have paid special attention to newer technologies—especially optical scan and direct record electronic (DRE) systems—because there seems to be general agreement that these two kinds of systems provide the major avenues for improvements in vote recording and tabulating systems.[5] There has also been some controversy over their relative performance.

The Cal-Tech-MIT study *Voting: What Is, What Could Be* (July, 2001)[6] concluded that optical scan systems and simple hand-count paper ballots had the lowest rates of residual voting in presidential, senatorial, and gubernatorial elections, while DRE machines lost nearly as many votes as punchcards and performed even worse down the ballot in gubernatorial and senatorial races. Lever machines performed well in presidential elections but did very badly in gubernatorial and senatorial races. The authors of the Cal-Tech MIT study concluded that "[optical] scanning is imperfect, but it is the best of what is," and they expressed surprise at the performance of DREs. They noted that electronic voting systems have "many apparent advantages" and that it is an "improving technology" with "great potential." Nevertheless, they concluded that "in terms of one very basic requirement—minimizing the number of lost votes—electronic voting does not have a very good track record. Paper systems have performed much better over the past dozen years."[7] We come to different conclusions.

We find that direct record electronic, lever machines, optical scan, and paper ballots <u>all</u> produce significantly fewer residual votes, between 1/2% to 1% less on average, than punchcards. This result remains true when the data are analyzed with a variety of statistical controls.

Using the four point scale devised by the National Commission on Federal Election Reform which uses the gradations of "good," "adequate," "worrying," and "unacceptable" performance, we find that all voting systems have a substantial number of unacceptable and worrying implementations, but punchcard systems are typically in the bottom two categories while all other voting systems—electronic, lever machines, optical scan and paper ballots—are typically in the top two categories.

We find, for example, that DREs perform as well as optical scan systems in the 100 largest counties in the United States which comprise about 40 percent of the

---

[2] Sequoia provided funding for research assistants, data collection, and supplies and expenses. The Principal Investigator (Professor Henry Brady) received no compensation for his services.

[3] As with all University studies, the SRC/IGS researchers undertook this study with complete freedom to do their own analysis and to form their own conclusions.

[4] We explain these terms in detail below.

[5] As we note below, counties have been moving away from paper ballots, punchcards, and lever machines. Paper ballots have found their niche in small, rural counties with an average population of 6000 people (98% of these counties have populations less than 19,000), but seem inappropriate for larger counties. Several studies, including this one, find that punchcards have excessive numbers of residual votes and that counties that move from punchcards to other systems almost always improve their performance. Lever machines are, as far as we know, no longer manufactured in the United States.

[6] The basic findings are on pages 19-23. More details are available in "Residual Votes Attributable to Technology: An Assessment of the Reliability of Existing Voting Equipment" at the website http://www.vote.caltech.edu.

[7] The first quotation is from page 22 of Cal-Tech/MIT (July 2001) and the remainder are from page 23. Although the authors of the report criticized the performance of DREs, they explicitly rejected the notion that "electronic voting is inherently flawed and should not be used (p 22)."

U.S. population and which include about one-third of the people using DREs for voting. We also find that DREs and paper ballots are less prone to what the Cal-Tech/MIT team call "very high residual vote rates (p 22)" than any other system. Optical scan systems do better than any other system when the simple average of presidential residual votes is considered for all counties, but this advantage disappears when we introduce statistical controls. In fact, lever machines seem to do best when controls are introduced with optical scan, DREs, and paper ballots tied for second place. Paper ballots do well in small counties.

Contrary to the Cal-Tech/MIT report's assertion that their conclusions "hold up to closer statistical scrutiny (p 22)," we find that the data and the evidence do not provide us with a clear-cut conclusion regarding the relative performance of any system except punchcards. All other voting systems—DREs, optical scan, levers, and paper sometimes perform well and sometimes perform poorly, but they always perform better than punchcards. Consequently, we do not come to any conclusions about the relative performance of non-punchcard systems. Moreover, we feel very sure that the available evidence does not provide any substantiation for Cal-Tech/MIT's claim that DREs might be harder to use than other systems. In fact, we present some evidence suggesting that DREs pose fewer problems for poorly educated voters than all other systems except paper ballots.

All in all, our claims are quite modest. We make modest claims based upon the available data because there are four major problems in assessing voting systems with national data:

Poor Data and Measures

The national data, as we document in detail later, are fraught with errors and problems including reports of more votes cast than voters who went to the polls and unreliable reports of voting system types. We found that data obtained from Election Data Services (EDS) had to be checked with information from individual states and even counties and that problems persist because of different reporting standards and conflicting information. Although we believe that the residual vote in the presidential race is a good measure of voting system performance, we also believe that residual votes for senatorial and gubernatorial races combine voting system performance and intentional undervotes to such a degree that it is very difficult, if not impossible, to separate out the two different phenomenon.

Difficulties Statistically Controlling for Other Factors that Affect Performance

Both the National Commission on Federal Election Reform chaired by ex-Presidents Carter and Ford (2001) and the recent report of the National Task Force on Election Reform (2001) comprised of election officials at the state and local levels contend that many factors affect the performance of voting systems.[8] County election administrations differ in their procedures for counting votes and in the reliability of these procedures. Voters vary in their ability to use voting machines and in their desire to vote in all political races. And voters' desires to record their votes depends upon the closeness and perceived importance of races. These factors are not constant across counties, or even precincts within counties, and they are not randomly distributed across different kinds of voting equipment. Consequently, differences in performance may be the result of differences in these conditions and not differences in technologies. Statistical methods that try to control for these differences must be used with great caution and substantial skepticism. Only when these methods converge on the same answer, as they do for punchcards, can we make a reliable inference.

[8] "The performance of voting systems is affected by several inputs that go beyond the equipment. Some of the most important are ballot design, voter education, and the skill and training of poll workers." (National Commission, August, 2001, p. 52) "To offer viable solutions, reform proposals must address problems and errors associated with 1) *People* (voters, poll workers, election administrators and staff, vendor personnel, candidates, and the media); 2) *Procedures* (vague and conflicting laws and inconsistent policies); and 3) *Technology* (outdated computer systems, voting equipment and tabulation systems)." (National Task Force, July, 2001, page 21).

Older Systems Remain in Situations Where they Work Best

Both the National Commission and the National Task Force have counseled that there may not be a system that is best for all jurisdictions,[9] and there is substantial evidence that some voting methods are especially well-suited to some locales but not to others. Paper ballots, for example, perform relatively well in the small rural counties where they are still used after having retreated in the past hundred years from major urban areas and large counties where they were prone to fraud and hard to count.[10] The good performance of paper ballot described by the Cal-Tech/MIT report must be put in this context—otherwise it would be easy to conclude from their report that paper ballots should be more widely used. But paper ballots have been around long enough so that a Darwinian selection process has limited them to the places where they perform well. It would be wrong to think that they could be transplanted elsewhere with similar success. The authors of the Cal-Tech/MIT report recognized this logic and wisely demurred from recommending a wider use of paper ballots despite the evidence for their high performance. But they apparently did not follow this logic to its conclusion by recognizing that selection processes have had more time to operate for older systems so that comparisons of old and new systems must take these processes into account.

Gross Categorizations of Diverse and Sometimes Rapidly Changing Technologies

A gross categorization of technologies into paper, lever, punchcard, optical scan, and electronic is necessary for statistical analysis, but it does not capture the nuances of these systems such as significantly different punchcard, optical scan, and electronic technologies and vendors with different records of performance. These problems may be especially acute for the newest and most rapidly evolving technology, electronic systems. These systems have only been implemented in the past twenty-five years, and one only has to think about how much personal computing has changed in that period to imagine how much the electronic systems of the late 1970s differ from those of the year 2000.

These four problems constitute substantial threats to producing valid conclusions about voting systems from the available data, and we strongly believe that the presence of so many threats to validity requires modest claims. Researchers must be sensitive to the many problems of making inferences about the relative performance of voting systems given the limitations of the available data, the problems of statistically controlling for differences in counties, the possibility that some systems have had the time to find their ecological niche and others have not, and the fact that characterizing technologies, especially rapidly evolving ones, is very difficult. We believe that the poor performance of punchcard systems[11] is clear even after these concerns are taken into account, but the relative performance of other systems cannot be determined.

Based upon our own analysis of data for the 2000 election and our concerns with these problems, we advise against the rush to judgment of the Cal-Tech/MIT report with respect to the relative performance of non-punchcard systems, and we agree with the Commission on Federal Electoral Reform and the National Task Force on Election Reform that much more testing of the available systems is needed before making any final judgments about the suitability of one system over another.[12]

[9] "Nor do we think one size will fit all...." (National Commission, 52) and "The committee does not believe that any one voting system or brand is at present suitable for recommendation for use in all jurisdictions." (National Task Force, p.35).

[10] See Joseph Harris, *Election Administration in the United States*, The Brookings Institution: Washington, D.C., 1934, for a description of the problems of paper ballots in major cities.

[11] Even for punchcard systems, we suggest that Datavote systems may perform much better than "Votomatic" style systems.

[12] Indeed, the Cal-Tech/MIT report calls for "significant investment by the federal government in research and development of voting equipment technologies and meaningful human testing of machines." (Cal-Tech/MIT 2001, p 3).

In addition, much more thought must be given to the characteristics of voting machines other than their ability to record and tabulate votes. Touch screen DREs that are like ATM machines, some of the earlier pushbutton DREs, and optical systems that scan ballots in a precinct have the advantage that voters can be given feedback that "checks their work" before they submit their ballot. Touch screen DREs have the potential for making it possible for language minorities to choose their ballot language in the voting booth and for those with disabilities to vote by listening to a recording. Optical scan, paper, and punchcards produce a paper trail that can be audited while lever machines and some electronic systems do not provide a way to check votes beyond a comparison of the counters or the electronic images in the machine with the voting lists. (Some electronic machines now have paper tapes that keep a record of votes.) In addition, voting systems differ in their costs, their portability and their ease of set-up. All these factors should be taken into account when choosing a system.

The remainder of this report documents our results. Section I on "Defining Performance" discusses the strengths and limitations of "residual votes" as a measure of voting system performance. Section II on "Types of Voting Systems and Their Susceptibility to Mistakes" describe the five systems studied in this report and their susceptibility to voter mistakes. Section III on "Data Collection Methodology" provides a detailed description of how we collected and cleaned the data used in the report. Section IV on "Statistical Analysis—Theoretical Issues" reviews a number of statistical problems that had to be addressed. These sections can be skipped by those who wish to get to the findings in Sections V through X.

Section V, "Descriptive Statistical Analysis for Presidential Residual Votes," reviews the performance of voting systems using a number of simple descriptive statistics to explore average and extreme performance. Section VI, "Multivariate Analysis for Presidential Residual Votes" adds statistical controls. Section VII on "Senatorial Residual Votes" presents information for senatorial races. Section VIII on "Presidential Residual Votes and Voter Education Levels" explores how lower education interacts with voting system type. Section IX presents some interesting "Findings about Specific Systems," and it shows that the gross categorizations of systems misses some important details. Section X takes advantage of California data to show how changes from punchcards to an optical scan system in Fresno County and from central count optical scan to an electronic system in Riverside County led to significant reductions in presidential residual votes. Section XI on "County Acquisition Decisions and Selection Bias" mentions an important confounding factor that must be considered when evaluating voting systems. Finally, the last section presents a "Summary of Findings and Recommendations."

I. DEFINING PERFORMANCE

Minimizing Residual Votes

Our report focuses narrowly on one dimension of performance: the degree to which voting systems record votes for every voter who comes to the polls. The events in Florida in 2000 provide ample reason for taking this criterion seriously, and other examples of close elections that might have been decided differently by uncounted votes are not hard to come by.[13]

Voting officials distinguish three major ways that discrepancies can arise between the number of people who come to the polls and the number of votes for all the candidates in a race. An undervote occurs for a particular office[14] when a voter comes to the polls and turns in a ballot but does not cast a vote for that office. An overvote occurs for a particular office when a voter marks more than one candidate for the office.[15] In these two cases, vote tabulation equipment should not record a vote for that office for that ballot. A "counting error" occurs when a properly marked ballot is not counted for that race.[16] The sum of undervotes, overvotes, and counting errors is the number of "residual votes."

Although a voter might make multiple marks and overvote as part of a protest or simply indifference, researchers and practitioners believe that most overvotes are the result of unintended mistakes.[17] The status of undervotes is more complicated. Some voters may *intentionally* decide not to vote in a race. Others may undervote *unintentionally* when they fail to make a mark that is counted by the voting system even though they tried to make such a mark. Intentional undervotes occur when voters do not want to vote for the candidates for some offices. Thus, there are typically more undervotes for less important offices (e.g., state assembly rather than President) or for matters farther down the ballot than for the top of the ballot. This "roll-off" phenomenon is well-known to election officials and political scientists (Burnham, 1965), and it can be quite considerable for matters not deemed very important by much of the population or in cases where ballots are very long. Unintentional undervotes occur when voters do not understand the proper way to make a mark to record their vote, when the voting system thwarts their efforts to make a proper mark, or when it simply does not pick up a mark that has been made (as with a hanging chad on a punchcard or too light a mark on an optical scan or paper ballot).

To get a sense of the size of these effects, consider the November 2000 elections in California where we have data on undervotes and overvotes for 46 of the state's 58 counties covering 83% of the voters. The overvote rate in the presidential election was about one-third of a percent (.35%) and the undervote rate was almost four times higher at 1.34% for a total residual vote rate of 1.69%. The overvote rate for Proposition 35, a relatively obscure matter regarding use of private contractors for public works, was .18% (less than that for the presidential race)[18] but the undervote rate was 11.16%.

Consequently, the residual vote of 11.34% for Proposition 35 was almost seven times greater than

---

[13] In the 1953 Italian elections, nearly 1.5 million spoiled ballots probably caused the Centre party coalition to fall short of a majority with just 49.85% of the vote (Carey and Carey 1957 p 201, 1958 p 570). Many newspaper articles and some magazine articles and technical reports have described how vote fraud or counting errors may have decided close elections (Analytic Systems 1974a Chapter 8, Saltman 1975 Chapters IV-V, Burnham 1985a, 1985b, 1986, Dugger 1988, Saltman 1988 Chapters 2,4).

[14] Throughout this report we talk about voting for an office, but the same arguments apply to voting for ballot propositions or other matters.

[15] The situation is more complex when the voter is asked to choose more than one candidate (e.g., three members for the city council) from a list. But in this case an overvote occurs when there are more marks than the number of candidates the voter is allowed to choose (e.g., four or more for the city council).

[16] A counting error can also occur if an undervote or an overvote is improperly counted as a vote. This probably sometimes occur, but the reverse problem—failing to count a properly marked ballot—is much more likely to occur. To take both these possibilities into account, our definition of "counting error" can be thought of as the net result of these two processes.

[17] The National Task Force on Election Reform (2001, p 22) states that "Overvotes are often due to voter confusion about the voting instructions but occasionally are intentional statements by voters."

[18] It seems likely that overvotes are more prevalent among those with less education. Such people are also less likely to vote for matters such as Proposition 35 that are less important and farther down the ballot. Consequently, the lower overvote rate for Proposition 35 may reflect the fact that those with less education were more likely to ignore the proposition altogether, thus making it impossible for them to make a mistake in voting.

for the presidency (1.69%) and the difference of 9.65% was almost entirely due to undervotes.

Because some voters roll-off and "undervote" for offices and propositions farther down the ticket, it has been argued that undervotes should not be considered a problem—they are simply a matter of voters choosing not to vote. For example, the National Task Force on Election Reform, composed entirely of elected and appointed state and local election officials, argues that "When examining undervoted ballots, the vast majority are clearly the result of voters intentionally skipping that contest. Rarely does an undervote mistakenly occur due to a voter incompletely marking or punching his/her ballot (page 22)." We agree that many undervotes for contests down the ticket (such as Proposition 35 discussed above) are the result of voters skipping the contest, but two kinds of evidence lead us to believe that the majority of reported undervotes for the *presidency* are unintentional. In a carefully executed paper using survey data, Knack and Kropf (2001) suggest that about .75% (three quarters of one percent) of those going to the polls intentionally decide not to vote. Knack and Kropf find that intentional undervoting is affected by the competitiveness of the presidential, senatorial, and gubernatorial races in a state and by the characteristics of the voters. Their figure of .75%, however, is substantially less than the approximately 2% residual votes for the presidency that we find in the national data. It seems likely that a majority of these residual votes are unintentional undervotes.

This result is corroborated by the substantial differences in undervotes in California counties in 2000 where the undervote rate for the presidential election varied from about one-third of one percent to over two percent in large counties.[20] (We focus on large counties because they are less affected by statistical fluctuations.) Similarly, in Florida counties in 2000, the undervote rate varied from one-eighth of one percent to almost three percent in large counties.[21] These within state differences might occur because counties have populations that differ in their proclivities towards undervoting, but they cannot be the result of differences in the competitiveness of the presidential and senatorial races within each state. Moreover, our analysis of California and Florida data shows that the type of voting system was also a factor in both states, even after using statistical methods to control for the characteristics of the voters. It seems likely, therefore, that the majority of undervotes for the presidential election are not intentional and constitute a failure of the voting system to record and tabulate votes properly.

Ideally, we would like to have information on both undervotes and overvotes to assess the performance of voting systems. We have that information for some states (e.g., most of the counties in California and Florida, and some counties in Texas), but it cannot be obtained for most states. Residual votes are a good proxy for evaluating how well voting systems record and tabulate presidential votes. Almost all overvotes and the majority of undervotes for presidential elections appear to be the result of unintentional errors which may be the result of voting systems that rob voters of their vote. Consequently, a residual vote measure for presidential elections, with proper statistical controls, can provide information about the performance of voting systems. We are much less sure about the adequacy of this measure for senatorial and gubernatorial elections because so many of the residual votes may be intentional undervotes whose numbers may differ substantially from one jurisdiction to another depending upon the closeness of the race.

Other Ways to Evaluate Voting Systems

Our report examines only one dimension along which voting systems can be evaluated. We believe that making sure that every person has an equal chance to have his or her vote counted is one of the major aspects of a voting system, but there are also other factors which should be taken into consideration when voting systems are designed and implemented. These factors include security of the voting process, system cost, maintenance and storage requirements, ease of installing the system and training poll workers in its use, the time and effort required to tabulate results, the ability to audit results, and so forth. We will have nothing to say about these factors, although it is clear to us that some of them have had and should have a great impact on the procurement decisions of election officials.

[19] Some of these people may overvote instead of undervote but it seems likely that most of them simply do not mark their ballots.

[20] For example, Fresno, Sacramento, San Bernardino, and Los Angeles are large counties with undervote rates of .30%, 1.39%, 1.67%, and 2.22% respectively. The variation is even larger in small counties.

[21] Brevard, Orange, Miami-Dade, Duval, and Sarasota are relatively large Florida counties which had at least 155,000 votes cast in 2000, and they had undervote rates of .13%, .35%, 1.44%, 1.89%, and 2.99% respectively. Washington and Hendry, smaller counties with fewer than 8,000 votes cast, reported undervote rates of 3.67% and 9.39%. Some other small counties reported undervote rates one-tenth to one-fiftieth this size.

II. TYPES OF VOTING SYSTEMS AND THEIR SUSCEPTIBILITY TO MISTAKES

Five Types of Systems[22]

There are five major types of voting systems used throughout the over 3,100 counties[23] in America: paper ballots, mechanical lever machines, punchcards, mark-sense (or optical scan), and direct record electronic (DRE). In the past forty years, there has been a steady move away from paper and mechanical systems, first to punchcards, but more recently from these three to optical and DRE systems. In most states, counties choose a single voting system for election day voting, but in some states this task is delegated to townships within the counties. As a result, a mixture of systems may be used within a county. Figure 1 shows the breakdown of counties using each kind of system, and Figure 2 shows the number of people using each kind. The two most widely used systems are punchcards and optical scan which together account for about 60% of the counties and the voters. More U.S. counties use optical scan (43%) systems than punchcards (17%), but more people vote using punchcards (32%) than using optical scan systems (28%). About 16% of the population use lever machines and about 12% use electronic systems. About 9% of U.S. counties use paper ballots but less than 1% of the voters live in these counties. The remaining 10% of the voters are in mixed counties, so we cannot identify which system these voters use.

Paper Ballots

Description The paper ballot is the oldest and the simplest method of voting, and it is now used in about 300, mostly small, rural counties or townships. A voter simply places a check next to the name of his or her preferred candidate on a piece of paper, and all such ballots are manually counted.

Sources of Mistakes This method of voting is familiar and quite transparent for almost everyone, but voters may make unintelligible marks—which can lead to undervotes or overvotes in the judgment of those counting them. Paper ballots are also sometimes paginated so that voters can miss a ballot page—leading to undervotes. Short of a visual inspection by the polling official which would violate voter privacy, there is no surefire way to check paper ballots for undervotes and overvotes, although voters themselves can easily check their work. In the counting stage, paper ballots must rely upon human beings who can easily make errors. Moreover, manual counting is difficult and time consuming when there are large numbers of votes.

Mechanical Voting Machines ("Lever Systems")

Description Lever systems were developed in the 1890's to facilitate voting and vote tabulation, and they are the oldest of the automated systems. They involve large displays of the entire ballot with small levers next to each choice. The voter typically enters the voting booth and closes a curtain by pulling a large lever. The voter then flips small levers next to the names of the candidates for whom he or she wishes to vote. After making all decisions and after a visual confirmation of those decisions, the voter then pulls the large lever, which counts each vote with counters in the back of the machine and which opens the curtain to allow the voter to exit from the voting booth. Mechanical voting machines are no longer manufactured, but they are heavily used in seven states (Connecticut, Georgia, Louisiana, New Jersey, New York, Pennsylvania, and Virginia) that account for about three-quarters of the approximately 400 counties where they are used.

Sources of Mistakes Difficulties can occur if the machine is improperly set-up or if the mechanism fails to relay vote choices to the counters. But overvotes are impossible when these machines are properly configured, and with such a simple interface with all races visible, unintentional under-votes are less likely.

[22] Our descriptions of voting systems draws from the published literature on voting systems (Analytic Systems 1974a,b; Moloney 1975; Saltman 1975; National Scientific 1977; Election Data Services 1981; Saltman 1988; Garber 1998; Florida Governor's Select Task Force 2001) and on voting system standards (National Clearinghouse of Election Administration 1984). It also draws upon numerous discussions with election officials, an examination of web sites that describe systems, and our own experiences with voting technologies.

[23] We will use the term counties throughout this report even though a small number of these jurisdictions, such as the Baltimore or St. Louis are cities and not counties.

Figure 1

Number and Percentage of Counties Using Voting Systems

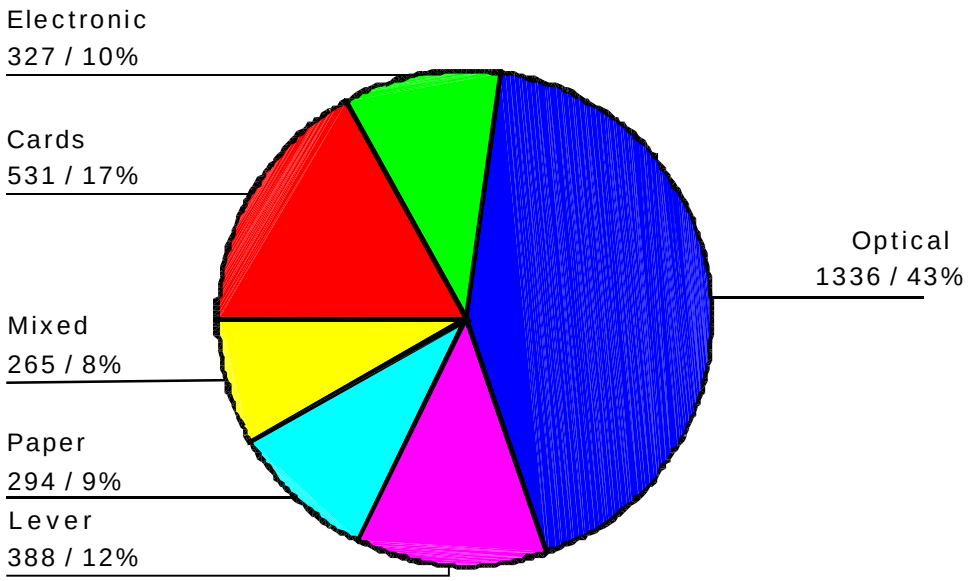
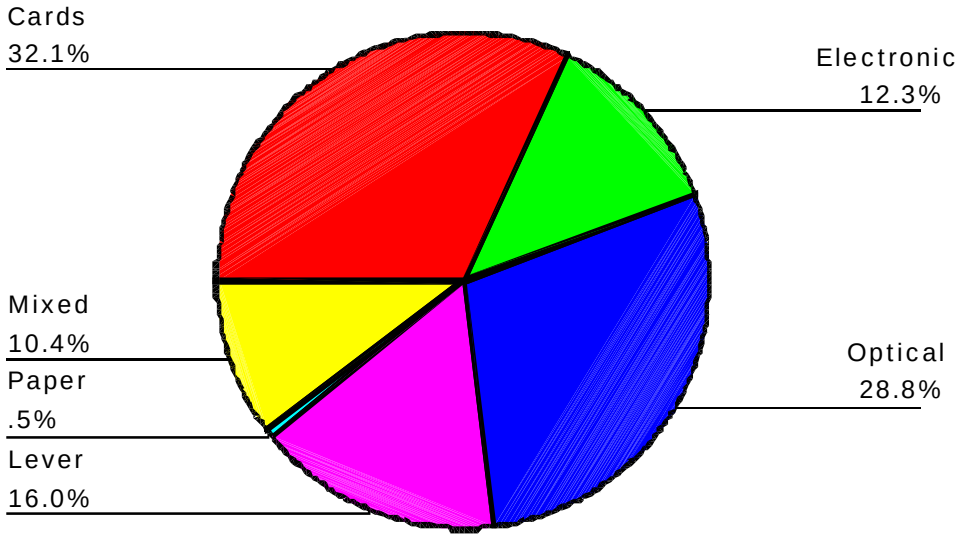


Figure 2

Percentage of Voters Using Voting Systems



Punchcard Systems

Description Punchcard systems were developed by Professor Joseph Harris of the University of California, Berkeley and his collaborators in the early 1960's. There are two variants of punch card systems. In the *Votomatic style* systems invented by Professor Harris, the voter receives a punch card with as many as several hundred small pre-scored perforations that can produce holes if the center of the perforated area is punched with a stylus. When punched out, the small pieces of paper are called "chads." The voter puts the punch card into a slot at the top of the Votomatic machine which has a ballot, typically consisting of a number of pages, attached to its face. When the punchcard is properly seated in the device, a series of holes in the Votomatic next to candidate names on the ballot are lined up with locations on the punchcard with the pre-scored holes. The voter then uses a stylus to punch through the hole in the Votomatic next to his or her choices. By punching with the stylus, the voter punches out the chads that correspond to the candidates for whom he or she wishes to vote.

In the *Datavote system* that was developed after the Votomatic, the voter receives one or more cards with the names of the candidates for various races on the cards themselves. The voter inserts the card into the Datavote machine and makes an immediately visible punch next to the names of the candidates for whom he or she wishes to vote using a mechanism that looks and performs like a one-hole punch.

Punchcard systems became very popular in the late 1960s because of their low cost and the speed with which ballots could be counted. Today, over 500 counties in more than thirty states use these systems, and in the 2000 election, there were eleven states scattered across the United States (Arizona, California, D.C., Florida, Illinois, Indiana, Missouri, Nevada, Ohio, Utah, and Washington) in which a third or more of the counties used them. Datavote style systems probably account for less than one-fifth of the counties using punchcards, although it is hard to be sure because many counties only identify their systems as "punchcards" without distinguishing between Datavote and Votomatic styles.

Sources of Mistakes Datavote systems retain a linkage between the mark that is made (through the punch) and candidate names, but Votomatic style systems sever this link. When the ballot is punched in the Votomatic machine, voters cannot see whether or not the chad has been effectively removed. If the chad is not removed, then an undervote can occur. To determine whether or not it has been removed, the voter must note a number on the Votomatic ballot that corresponds to the candidate that is chosen, take the punchcard out of the machine, and then check to see whether the hole with the number has been punched through. Moreover, if the voter wants to be sure not to overvote, he or she must note all of the numbers for that office, and make sure that only one is punched through corresponding to the chosen candidate. Problems can thus occur with the Votomatic because chads are inadvertently punched out or because not enough pressure is used to punch the chad out completely. A related source of error with Votomatic style machines is that the punchcard may not be properly seated so that candidate names and holes on the punchcards will not correspond to one another.

Because Datavotes often involve multiple cards and because Votomatics often involve multiple ballot pages, undervotes can also occur because of missed cards or pages. Finally, neither system has traditionally involved any kind of automated checking for undervotes and overvotes, although a recent implementation of a Votomatic style system in Chicago provides for that possibility.[24]

[24] Cook County, in which Chicago is located, had a 6.18% rate of residual votes in 2000. We classify this system with all other punchcard systems.

Mark-sense or Optical Scan Systems[25]

Description Optical scan systems use paper ballots, but the mark that is made by the voter is recorded using special sensing equipment. Although this technology goes back almost half a century, mark-sense machines began to be practical in the 1980s and 1990s. In these systems, voters fill in bubbles or make some other kind of prescribed mark on the ballot with a pencil or pencil-like device. This mark is then scanned by a machine to record the vote. Scanning can be done centrally or in each precinct, although precinct scanning is a more recent innovation. Centrally-scanned ballots are taken to a small number of locations within a county to be scanned together, but precinct scanning is done where the voter votes. Over 1300 counties in more than forty states use optical scan systems, and these counties are about evenly split between those using central and precinct scanning systems. Precinct scanning systems, however, are replacing central scanning systems in many areas.

Sources of Mistakes Optical scan devices require that the mark be sensed (hence the name "mark-sense") by an infra-red or some other kind of scanner. Typically this requires that a specific kind of marking device (such as a "number two" pencil) be used. If the wrong device is employed, if the mark is not substantial enough, or if stray marks are on the page, then undervotes or overvotes can be incorrectly recorded. Because the ballots are pieces of paper, they can also be mislaid or damaged.

Precinct scanning helps to alleviate these problems. Votes are recorded immediately so that a voter can be sure that it will not get lost, misplaced, or damaged when the paper forms are transported to a central location. Probably more importantly, precinct scan systems typically allow voters to have their work checked by the machine. In the most advanced implementations, the precinct scanner can report undervotes and overvotes as well as ballots which appear to be completely blank. While not all precincts necessarily utilize these features, it seems likely that they can help to catch some errors.

Direct Record Electronic Systems

Description Electronic systems are the newest voting systems, and they are now used in more than 300 counties around the country. There are two primary distinctions for DRE Systems. The first is between the somewhat older *push-button* and the newer *touch screen* systems. A push-button system is like a lever machine, except that it has buttons that electronically record votes instead of levers that turn counters in the back of a machine. There are different versions of push-button systems, but in many of them, a light will come on next to candidates names when they have been selected—thus providing voters with feedback indicating that their choice is recognized by the system. Touch screen systems are exactly what the name implies. The names of the candidates appear on a screen, and voters touch the names for whom they wish to vote. At the end of their session, they have the option to review their votes to make sure they have cast them properly.

The other important distinction for DREs is between *paginated* and *full-faced* systems. All touch screen systems are paginated, and at least one push-button system—Microvote—is paginated although most are full-faced. A paginated system is one in which each screen contains options for a limited number of races, and when the voter has finished with the races available on the page, he or she moves on to the next page, until the end when there is typically a chance to review all votes. A full-faced system is one in which there are buttons for every race on a single machine face, with names of candidates next to each one. The voter simply presses the button for a candidate in each race, and hits a "vote" button when finished, much like a lever system. Full face systems make it harder for voters to miss whole blocks of races.

DRE systems also differ in the way they handle straight party line voting which is a state mandated option in 17 states. In straight party line voting, a single vote can be cast for a party indicating a preference for all candidates of that party rather than having to mark each candidate individually. DRE systems can handle this option in two ways. The first method begins by asking voters if they wish to vote a straight party ticket. If they do, they may then review

[25] "Mark-sense" is the more general category because the mark that is made may not be optically sensed but may use some other wavelength or method of sensing. Most mark-sense devices, however, are essentially optical scan systems.

the individual offices, but in order to change their votes for a specific office, they must first deselect the party line option, and then they can change their votes for individual offices. The second method allows voters to review the offices after choosing a party line ticket and to change votes wherever they choose without first deselecting the party line option. Careful programming is especially important with these systems.

Finally, there is still another feature of DREs in states with straight party-line voting. DREs come with two options for dealing with party-line votes. They may be set so that a party line vote includes votes for all offices, or they may be set so that a party-line vote includes votes for all offices *except* the president, and presidential votes must still be cast separately. If the latter option is selected, proper voter notification in the program is extremely important.

Sources of Mistakes DRE systems, especially touch screens, use technologies that may be unfamiliar to less well-educated or older voters. As a result, special attention must be paid to making sure that their interfaces are easy to use and that voters fully understand their options. Paginated systems may lead to undervotes if voters accidentally skip a page and do not choose to carefully review their ballots at the end of their voting session. Errors in configuring the systems by election officials or others can also lead to difficulties. But DRE systems, like mechanical lever machines, do not allow overvotes, and the touch screen systems can provide the voter with a review of all votes to insure correctness and to reduce the possibility of undervotes. DREs also present some advantages for foreign language speakers and those with disabilities.

Absentee Ballots

The substantial and growing number of absentee ballots present some complexities for our analysis. Residual vote information is generally not available separately for them, and absentee voting may be quite different from election day voting. In 1996, the proportion of absentee voting ranged from about two to three percent of total votes (for almost twenty states) to twelve percent or more (for six states). In 1996, 46% of Oregon's vote and 37% of Washington's vote was absentee. In 2000, Oregon went entirely to "vote-by-mail."[26]

The differences between absentee voting and election day voting can affect the evaluation of voting technologies in two ways. First, voting by mail is clearly a different experience than voting in a precinct because there is more time to ponder, but less help immediately available from poll workers. Therefore, even if the same system is used for voting by mail as voting on election day, there are probably differences in the kinds of errors that can occur.

But to complicate matters, the same system is not always used. In the case of mechanical lever and DRE systems, the reason is obvious. Voters cannot be sent these machines in the mail, although they can be asked to vote early at a precinct or at an office using one of these systems. In fact, most counties using DRE and lever systems use some other method for absentee voters such as paper ballots, optical scan ballots, or punchcards. For paper, punchcards, and optical-scan systems, voters can be sent ballots that are similar to those used by election day voters, and many counties do this. For punchcards this sometimes means sending out a punchcard with instructions about how to manually punch-out chads. The virtue of this procedure is that it might make the Votomatic style system more transparent to an absentee voter since chads have to be punched out directly, but it might also lead to damaging the punchcard in one way or another. For optical scan systems the voter may not use the proper marking device or fill-in the form correctly. In addition, the benefits of precinct scanning cannot be enjoyed so voters do not have their work checked. Consequently, some counties simply treat the ballot as a paper ballot and count it manually.

To the extent that absentee voting is a different experience from election day voting, our results regarding voting systems may be affected. The consequences for residual vote estimates of having different systems for election-day and absentee voting can be calculated by noting that at one extreme, the absentee systems cannot do better than zero overvotes and zero undervotes, and at the other extreme they probably cannot do too much worse than about twice as bad as the election day system. For most states, the impact cannot be too large because the percentage of absentees was less than twelve percent in 42 states in 1996. Consequently, since the average election day system has about 2% residual votes, the maximum

[26] Our thanks to our colleague Raymond Wolfinger who supplied these figures taken from the 1996 Voter Supplement to the Current Population Survey.

effect that an absentee system could have in these states would be to reduce this figure to 1.8% or to increase it to 2.2%.[27] But for some states, or counties within states, this effect might be even bigger. As a result, we should keep in mind that we should look for larger differences than these amounts if we are to say that one system is better than another.

In addition, it is worth noting that absentee votes may have a greater impact on our estimates of performance for mechanical lever and DRE systems because they typically use other methods for absentee votes. Indeed, for Riverside County, California, for which we have detailed data, 37% of the voters voted absentee and the residual vote percentage for absentees who used an optical scan method was about double that for the DRE system that was used on election day.

Absentee votes also raise another difficulty. There may be some counties where there are negative residual votes because the count of the total number of ballots cast does not include the absentee ballots. We do not think this is a big problem, but it remains a possibility given the inconsistent reporting standards for total turnout.

Another growing trend, early voting, might also affect our results, but it was less than or equal to 3% of votes cast in all but eight states in 1996. Nevertheless, there are some states (e.g., Texas) in which it constitutes a substantial fraction of the voters. Typically, however, early voting often uses the same system as election day voting so that it probably creates less bias than absentee voting.

System Features and the Sources of Residual Votes

The preceding discussion identifies five features of voting systems that can produce unintentional residual votes:

1. *Whether or not the names of the candidates are attached to the marking place.* Paper, lever, optical scan, Datavote punchcard (but not Votomatic style), and DREs all have a strong tie between the place where the voter marks the ballot (by writing, punching, pushing, or touching) and the name of the candidate that is chosen. As noted above, Votomatic style punchcards sever this link.

2. *The certainty of the method of marking.* All systems can fail because of defects in the marking process. Paper can fail because marks are unintelligible. Punchcard systems can fail because a thorough punch has not been made. Optical scan can fail because the voter uses the wrong marking device or does not use it correctly. Lever machines and DREs can fail if there is some problem with the connection between the voter's pushing of a lever, a button, or a screen, and the actual recording of a vote. As noted below, however, precinct scanning machines for optical systems or DREs can be programmed to provide feedback about possible mistakes to the voter.

3. *Full-face versus pagination.* Lever machines and all but one push-button DRE are full-faced systems that present the voter with the entire ballot all at once, thus reducing the possibility of missing some races, although the sheer size of the display might create some confusion for voters. All other systems are typically paginated to some degree or other depending upon the number of races and propositions on the ballot. In the case of some optical scan ballots, the ballot may be made very large to avoid people missing part of the ballot, but there is a limit to the size that can be handed to a voter and some counties use the reverse side of a ballot as well as the front leading to the possibility that a voter will overlook one side or the other.

4. *Checks for overvotes or undervotes.* Paper and punchcards (with the exception of the Cook County system described above) do not allow for automatic checking of overvotes or undervotes. Lever machines and DREs are programmed to make overvotes impossible. Undervotes can be reported to voters in precincts with optical scan systems and on touch screen DREs right in the voter's booth, but some counties are reluctant to

[27] The calculation is simple. For example, if the election day system has 2% residual votes and the absentee system involves 10% of the voters with no residual votes, then the observed residual will be [.90*.02 + .10*.00 = 1.8%]. If the absentee system has 4% residual votes, then we obtain [.90*.02 + .10*.04 = 2.2%].

enable these features because they fear it will slow-up voting. In addition, precinct optical scan systems typically require voters to leave the voting booth to check their ballot, leading to issues of voter privacy.

5. *Method of counting.* The method of counting can produce difficulties in a number of ways. Manual counts of paper ballots can produce undervotes, overvotes, and even too many votes given the number of ballots cast. Mechanical lever machines require reading counters at the back of the machine. These numbers must then be recorded, and finally reported to the county board of elections. Errors can occur at each step of this process. Punchcards and central optical scan systems can suffer from lost or damaged ballots and from the need to "remake" a ballot by removing chads or by marking over unreadable marks made by voters. Precinct optical scan and DREs can ensure that ballots are counted, but those vote counts must then be properly reported to the elections officials.

Many factors can cause overvotes and undervotes. When used carefully by voters and maintained properly by election officials, all voting systems should produce very few mistakes, but voting systems are not used by experts and the quality of maintenance varies. Voting systems are used by voters who encounter them infrequently and who have little or no training in their use. Precinct poll workers typically have no more than a few hours of training in which the operation of voting systems is only one of many topics. Counties vary tremendously in their ability to set-up and maintain voting systems. Thus, we would expect that residual vote rates would be affected by the characteristics of the voters in a county (especially their educational level) and by the characteristics of the county administration as well as by the characteristics of the systems themselves. In effect, each system is a vote recording and tabulating technology that requires inputs of capital (the system itself) and labor (provided by the voter, the poll workers, and election officials) to produce a good outcome. Our methodology in this study has been designed to take these facts into account.

III. DATA COLLECTION METHODOLOGY

The objective of this project was to determine the effect of different voting systems on voter mistakes. It would be ideal for this purpose to have information on both undervotes and overvotes. Unfortunately, very few counties and states keep track of overvotes and undervotes separately (indeed, some states and counties keep track of neither), so our study focuses on "residual votes." We define the residual vote for each county as the total number of people who showed up to the polls to vote minus the total number of votes cast in the presidential election. We also define a similar figure for senatorial and gubernatorial races. The presidential residual vote serves as our best measure of residual votes across the country. Unlike any other political office, the presidential race provides a constant reference point because every eligible voter is given a chance to vote on the presidency with the same major candidates and many of the same minor candidates.[28] As a result, residual votes for the presidency are mostly unintentional overvotes and undervotes which can be used to measure the performance of voting systems. A voting system which is highly reliable and unlikely to lead to voter error will produce a low residual vote (although there will always be a few people who simply did not want to vote in the presidential race), and a voting system that is difficult to use or mistake-prone will produce a high residual vote.

Residual votes for senatorial and gubernatorial races are much more problematic because the closeness and importance of these races has a lot to do the number of undervotes, thus making the residual vote measure a reflection of not only voting system performance but also voter interest in the race. Indeed, there are some states in which the race for Governor, Senate, or even some other office out-polled the presidential race, and we will show that the competitiveness of these races has a discernible impact on the number of residual votes.

We focus on residual votes in counties in the 2000 general election. We consider counties because they are the units that, in most cases, manage elections most directly and determine voting systems. Most of our emphasis is on the presidential race because we are more skeptical about what residual votes mean for races down the ticket. We have used data for 2000 because it the only information of suitable quality and, luckily, it also provides the most up-to-date picture of the relative performance of voting systems.

Data Requirements

Our approach requires three different types of data. We need to know the specific voting system each county used in the 2000 general election. We need the total votes cast and total ballots cast in each county for each race of interest in the 2000 general election in order to compute residual votes. We need information on voters, counties, and political races to control for the characteristics of the voting age population, the counties, and the competitiveness of the race that might affect the level of residual votes. For example, less educated voters may have more trouble understanding how voting systems work and thus be more likely to produce undervotes or overvotes, and smaller counties might be more variable in their election administration so that they would sometimes produce substantial numbers of undervotes or overvotes.

We began our data collection by purchasing a data set from Election Data Services (EDS). This data set contained election data and voting system data by county for the 2000 election. Before proceeding with our analysis, though, we thoroughly audited the data set to insure that we had the best information available since analysis conducted on incorrect data would be futile. As described below, we have spent most of our time and effort trying to correct these data which had many errors.

Voting System Data

The voting system data reported to us by EDS had numerous errors. Our first check was to compare the voting systems EDS reported for each county in the 1996 to the 2000 elections, and we found many irregularities that made no sense, such as a very large number of counties moving from newer and

[28] States differ in the conditions required for presidential candidates to get on their ballots so that in 2000 there were from four to ten candidates listed on presidential ballots, but every state listed the two major candidates. Even so, there is some evidence (Knack and Kropf, 2001) that differences in the competitiveness of the presidential race affect intentional undervoting, and we have added a measure of competitiveness to our multivariate results.

presumably more reliable systems such as optical scan systems to punch card systems. We investigated further and discovered that substantial portions of the data from EDS conflicted with election reports from Secretaries of State (SoS). For these reasons, we contacted the SoS for every state in the country to get their report on what system each county used. For some states, this was a matter of finding reports on the SoS's website. For those states that did not list county voting systems on their websites, we called their offices for copies of their reports. We then compared the system that EDS reported to the system that the SoS's reported for every county in the country. Whenever both agreed for every county in a state, we assumed that we had the correct information about those counties. There were 23 such states: Alabama, Alaska, Arizona, Arkansas, California, Connecticut, Delaware, DC, Florida, Hawaii, Idaho, Iowa, Maryland, Missouri, Montana, Nevada, New York, Oklahoma, Rhode Island, South Carolina, South Dakota, Washington, and West Virginia.

In 17 states, the SoS and EDS agreed on the majority of counties, but there were some discrepancies. For every county for which we found a discrepancy between what EDS reported and what the SoS reported, we called the county election board and asked them what system they used. Generally, they agreed with the SoS, and we accepted those as correct, but in the few cases in which they did not agree, we assumed that the county was correct since they deal more closely with the matter. Those states were: Georgia, Illinois, Kentucky, Louisiana, Mississippi, Minnesota, Nebraska, New Jersey, New Mexico, North Carolina, North Dakota, Ohio, Oregon, Pennsylvania, Utah, Virginia, and Wyoming.

For 5 states, there were so many discrepancies that we could not call every county. For such states, we called from one third to one half of all discrepant counties, found that the SoS agreed with the counties, and accepted the SoS's word on the remaining counties. Those states were: Colorado, Indiana, Kansas, Tennessee, and Texas.

There were six states for which elections were run by townships, not by counties, so that nearly all counties had mixed systems (some areas within the county used, for example, optical scan while others used punch cards). Since we cannot analyze mixed counties with county data, we ignored them in this analysis. Those states are the four states of upper New England: Maine, Massachusetts, New Hampshire, and Vermont and two states of the upper midwest: Michigan, and Wisconsin. This decision left us with 44 states plus the District of Columbia for analysis.

We are relatively confident in our classifications of county voting systems, and for every county included in the data, we have at least two sources agreeing on the system used, but we still believe that there are unavoidable errors given the difficulty of data collection and the erratic record-keeping of some counties.

Although we obtained data for 1996 from EDS, we did not use it for several reasons. After finding many errors in the 2000 data, we became skeptical of the quality of the 1996 data. We have no reason to believe that these data would be any more reliable than the 2000 information which was quite error ridden. We could have checked it in the same way that we checked the 2000 data, but we decided that too many states and counties would be unable to provide reliable data in 2001 about their 1996 election systems. In short, auditing the 1996 data with the same care with which we audited the 2000 data would be impossible. It is unfortunate that we do not feel confident about these data because they would be useful for statistical analysis, but our firm judgment is that any attempt to use them would be so riddled with error that any conclusions based on them would be highly suspect at best.

Election Data

Two types of election data are needed to compute residual votes: total ballots and total votes cast for presidential, senatorial, and gubernatorial candidates. EDS provided county-level returns for almost all states for these races, showing total votes cast for all candidates. These data were considerably more reliable than the voting system data. Four states were completely audited, and EDS figures perfectly agreed with SoS numbers. Also, we selected approximately 100 counties at random and compared EDS figures to the numbers reported by the SoS. While not all of those numbers were a perfect match, discrepancies were numerically small (rarely more than double digits when total votes were in the thousands), and

probably occurred as a result of some counties finishing their final counts of absentee ballots late, or data entry errors. Since the discrepancies are mathematically insignificant, we used EDS data as given for total votes cast.

Unfortunately, EDS did not report total ballots cast in the counties for the following eighteen states: Alabama, Alaska, Arkansas, Colorado, Delaware, Georgia, Indiana, Maine, Mississippi, Missouri, North Carolina, Oklahoma, Pennsylvania, Tennessee, Texas, Virginia, West Virginia, and Wisconsin. Since virtually all county voting systems were mixed in Maine and Wisconsin, we did not collect further data for these two states, but for all of the other states just listed, we contacted the SoS and requested ballots cast or turnout by county. The following nine states were able to provide such data: Alabama, Colorado, Delaware, Georgia, Indiana, North Carolina, Texas (numerous counties were missing in Texas, but many counties did report residual votes), Virginia, and West Virginia. Counties in the remaining seven states (Alaska, Arkansas, Mississippi, Missouri, Oklahoma, Pennsylvania and Tennessee) had to be excluded from our analysis because we could not calculate residual votes. Additionally, we collected county-level data on votes cast from the SoS of Georgia and Virginia, for whom EDS provided ballots cast, but not votes cast.

The quality of these data bears some discussion. Total votes is a much more important figure than total turnout because it decides who wins and loses elections. Even so, the number of total votes often changes in the days and weeks after an election as votes are tabulated again, provisional votes are checked and added to the total, absentee ballots are counted, and the overall results are audited.

Total turnout is also reported, but it is a much less important figure to election officials than total votes leading to the possibility that even less care is given to determining this figure. In fact, even the definition is somewhat uncertain because some states report (or attempt to report) total turnout—the number of people coming to the polls, and other states report total ballots—the total number of ballots involved in the vote counting process. Turnout can be larger than total ballots if entire ballots are lost in the counting process or if some voters fail to register their votes by not turning in their ballots or by failing to execute the action that casts their vote on a voting machine. Furthermore, different methods are used to measure turnout than total ballots. Turnout is typically measured by counting the number of people who signed register books while total ballots is measured by counting the total number of people who turned in ballots by submitting a paper record or by pushing the right lever or button on voting machines. Many errors can occur in these counting processes. Some states scan register books which can lead to errors from illegibility, crossed-out information, or the imperfections of the scanning software. Miscounting can occur because absentees or provisional ballots are not included in final counts of total ballots. We found many cases where election officials explained their record of negative residual as the result of one or more of these kinds of problems. The net result is that we have a substantial skepticism for the turnout data that we have collected. Moreover, we worry that there may be a complicated interaction between state laws for reporting turnout and voting systems that may affect the total ballot/turnout figures that we use to compute residual vote.[29]

Demographic and Political Data

Finally, for some of our analysis, we investigate whether or not differences between residual votes in different counties result from voting systems or characteristics of the counties or their populations. In order to provide some statistical controls for these factors, we used county-level data on demographics from the 1990 census because the same breakdowns from the 2000 census have not yet been made available. We are as confident in the accuracy of these data as we can be in any data set, except that the numbers are for 10 years prior. We assume that the changes over the last 10 years in county demographics do not systematically bias our results.

We also used standard competitiveness ratings for the presidential, senatorial, and gubernatorial races. The senatorial competitiveness ratings are based on the classifications of the *Cook Political Report* and the *Rothenburg Political Report* for the month preceding each election, with noncompetitive races coded as 0 and "toss-ups" coded as 1, with in-between gradations being coded as 1/3 or 2/3. For the Senate races, the

[29] For example, officials in some Pennsylvania counties explained that they had more votes than ballots because provisional votes were added to the vote total but not to the ballot total. Officials in Kentucky explained that the scanning of registers to determine turnout often led to figures that were did not jibe with the total number of votes. We also encountered 17 cases with exactly zero residual votes and another score of cases where the small number of residual votes seemed incredible to us.

competitiveness measure is the average of the Cook and Rothenberg scores. The gubernatorial measure is from the *Cook Political Report* for the month preceding the election, and it is scored in the same way as the senatorial measures. Presidential competitiveness is measured by the *Cook Political Report's* score using the same scale as for the senatorial races. These measures may tend to overstate the actual competitiveness of many of these races. We believe them to be fairly good proxies for the perceptions in the public of the competitiveness of the races, because much of both the print and television media rely on the same data sources as these reports, if not on the reports themselves.

The Final Data Set

The final sample consists of 38 states. The six township states of upper New England (Maine, Massachusetts, New Hampshire, and Vermont) and the upper midwest (Michigan and Wisconsin) had to be excluded because the townships used different voting systems so that almost all counties had a mixture of systems. Five southern states centered on the Ozark Mountains (Arkansas, Mississippi, Missouri, Oklahoma, and Tennessee) had to be excluded because there was no data on the total number of ballots cast. Alaska and Pennsylvania had to be excluded for the same reason. The resulting sample of 38 states consists of 2396 counties.

A few more exclusions were necessary because of missing or clearly erroneous data for various counties. Twenty-four counties were eliminated because the voting systems were unknown or mixed (two in Kentucky, eleven in Minnesota, one in Montana, seven in Texas, two in Virginia, and one in Wyoming). One hundred eight counties were eliminated because there was no information about the total number of votes cast (106 in Texas, one in Alabama, and one in Hawaii).[30] The only state that was seriously affected by these exclusions was Texas for which we have information on only 141 of the 254 counties—although we only lose 31.5% of the population. Finally, we excluded three small counties (two in Alabama and one in Nevada) because they had residual vote rates of 61%, 49%, and 23%, and forty-two counties with negative residual vote rates. Some of the reasons why we obtained negative vote rates are described above and in the Appendix which also lists the states with these rates. The resulting data set of 2219 counties constitutes almost three-quarters of the counties, three-quarters of the states, three-quarters of the population, and three-quarters of the presidential votes cast in the 2000 presidential election. Thus, our data set provides a very representative sample of the voting systems and voting data across the country.[31]

Comparison with the Cal-Tech/MIT Data

Since our results differ from those of the Cal-Tech/MIT study in some important ways, it would be useful to compare our data with theirs. We have requested an exchange of data so that we could do this in detail, but the Cal-Tech/MIT team has indicated that they are still finalizing their data so that they cannot currently provide us with their data in exchange for ours. Thus, our comparison has to rely upon the limited information provided in their reports.

The biggest difference between the Cal-Tech/MIT report and our report is that (with a small exception noted below) we do not use residual vote data for elections before 2000.[32] We do not believe that these data are reliable, and we believe that it is a mistake to use data that are likely to be so error-prone. Yet, while we do not use these earlier data, we have spent a great of time and effort obtaining data from the 2000 election. Indeed, it appears that we have data for some states that the Cal-Tech/MIT team does not have (Alabama, Colorado, Delaware, Minnesota, and West

[30] The (very small) county in Hawaii had its votes lumped together with those of another county.

[31] The data set still has a few odd numbers in it, but their inclusion or exclusion does not significantly affect the results. Some of these oddities are discussed in Appendix 1.

[32] As a result, we cannot use the time-series cross-sectional methods that they employ. These methods can be very useful, but we do not believe that poor data should be used just because good methods are available. Unfortunately, good methods seldom, if ever, repair the damage stemming from poor data, and time-series cross-sectional methods are not specifically designed to overcome poor data quality.

[33] We infer this because on page 89 of their report, the Cal-Tech/MIT team lists fourteen states and the District of Columbia for which they do not have enough data to report a residual vote. We have information on five of these states. The first footnote in the March 30, 2001 paper of the Cal-Tech/MIT Voting Technology Project suggests that they have data on two other states (Massachusetts and Vermont) for which we do not have detailed information about voting equipment.

Virginia)[33] and the District of Columbia, but it is hard to be sure given the information they provide. For the 35 states for which they report residual vote rates for the 2000 election on page 89 of their report, we have information on 34 of them. Our computation of the average residual vote rate is identical to theirs for 23 of these states and within a practically insignificant plus or minus .1% for 8 others. There are significant differences for three states: Indiana, Maryland, and Nevada.[34] We found some significant errors in Indiana which lowered its average, but we are not sure why there are discrepancies in the other two states. We do know that we have spent many hours cleaning our data to make it as reliable as possible, and we have provided a detailed account of the ways that we have done this.

[33] We infer this because on page 89 of their report, the Cal-Tech/MIT team lists fourteen states and the District of Columbia for which they do not have enough data to report a residual vote. We have information on five of these states. The first footnote in the March 30, 2001 paper of the Cal-Tech/MIT Voting Technology Project suggests that they have data on two other states (Massachusetts and Vermont) for which we do not have detailed information about voting equipment.

[34] The figures are: Indiana (1.2% for us, 2.1% for them); Maryland (.8%, .5%) and Nevada (.7%, 1.3%).

IV. STATISTICAL ANALYSIS—THEORETICAL ISSUES

The Pitfalls of Comparing Average Residual Votes for Voting Systems

The simplest and best way to evaluate voting systems would be to conduct a randomized experiment in which voting systems are randomly assigned to counties so that a random selection of counties employed each type of voting system. This approach would insure that on average the characteristics of the counties for each type of system would be the same. As a result, the average residual votes for one system could be fairly compared with those of another and standard statistical tests could be used to see if there were significant differences.

When simple averages are used to characterize voting systems in the United States, as in some of our analysis below or in the Cal-Tech Report (July, 2001, 21), researchers are acting as if the allocation of voting systems is essentially random and counties using each kind of system are, on average, similar. Yet a glance at a map of the distribution of systems, or an analysis of county characteristics by system type, or some thought about the ways in which systems are procured quickly discloses that this is not so. Voting systems are far from being randomly assigned to counties. Consider just the differences in county characteristics across system types. The average population of counties using paper ballots is one-tenth that for the counties using any other kind of system. Median family income is $2000 higher in those counties using punchcards than in those using any other kind of system. The percentage of minorities is 50% higher in counties using levers than in those using any other kind of system. The percentage of those going to high school is five percentage points lower than the national average in counties using DREs or lever machines.[35] The competitiveness of Senate races is much higher in counties using lever machines or paper ballots than those using other systems. Because these differences undoubtedly affect residual vote rates, simple averages must be used with great caution.

But even if averages could be compared fairly, an average is not the only statistic that should be used to describe the performance of voting systems. Averages do not tell us much about whether a system is prone to very poor performance. Although an election official who is looking for a new system might overlook a few instances of poor performance on the grounds that even excellent systems sometimes perform badly because of poor implementation, the official would also know that many examples of bad performance probably signify some problem with the system itself because so many counties seem unable to implement the system effectively. The official would also know that the goal of a good administrator is to perform better than the national average so that it makes no sense to procure a system that always performs at the average but seldom performs appreciably better than that. These considerations show that the average is an incomplete measure of system performance. It is also important to know something about the best and worst performance of a system.

How can this be determined? Rather than using unfamiliar statistical measures, we believe that the easiest and best approach is to use the grading system presented in the report of the National Commission on Federal Election Reform. On page 55 of their report, the Commission presents a four point scale of performance based upon residual votes:

Good:	Zero to 1%
Adequate:	1-2%
Worrying:	2-3%
Unacceptable:	Above 3%.

We use this scale throughout this report and we pay special attention to "Good" and "Unacceptable" performance. We also follow the lead of the Cal-Tech/MIT report in which they characterized residual vote rates of 5% and above as "very high residual vote rates" (July, 2001, p 22).

[35] These are unweighted averages. To make things even more complicated, the population weighted averages tell a somewhat different story.

System Performance May Vary by County Characteristics

A knowledgeable public official would also know that the performance of systems differs depending upon the county's size and complexity and the characteristics of the people who vote in the county. A system, such as a paper ballots, that might work very well in small rural counties with geographically large but sparsely populated areas, might work very badly in densely populated urban areas. Or a system that might work well with a young population might not do so well with an older population. Therefore the official would want a system that performed well in counties like hers.

This consideration suggests that even in a randomized experiment, the relative performance of voting systems would differ according to the characteristics of the counties. We suspect, for example, that in such an experiment, paper ballots would do very well for small rural counties but very badly for large urban ones. On average, paper ballots might appear to perform well, but it would be an egregious mistake for a large urban county to implement paper ballots based upon this average performance.

Consequently, the performance of systems must be broken down by county characteristics. There are many such characteristics, but we think that the most obvious one is the population of the county. Large counties are typically much more urban and more complex than smaller counties. They have hundreds of precincts. They also have larger staffs of election officials. The heterogeneity and complexity of these counties suggests that they might have large residual voting rates, but their larger staffs of election officials suggest that they might be better able to cope with the problems of election administration. Figure 3 tells the story that comes from our data. The presidential residual vote rate goes down significantly from about 2.7% in smaller counties to 1.6% in larger counties. We do not know what mechanism causes this to be so, but it is an important descriptive fact about presidential residual votes and the size of the county.[36] It provides another reason to consider large counties separately, and we will do so. We do not claim that this is the only way to think about subgroups of counties, and we certainly do not want to claim that it is the best way. But we do think it is a reasonable way. Indeed, the National Commission the Federal Election Reform saw fit to consider explicitly the forty most populous election jurisdictions in the United States. We will follow their lead and consider both these forty areas and the top 100 areas. The top 100 areas comprise all counties with populations of more than approximately 500,000 people. The total population in these counties is over 100 million—about 43% of the United States population.

How Many Units?

Our decision to consider the 100 most populous election jurisdictions as well as all counties in the United States brings up an interesting question: How many units are we analyzing in either case?[37] In statistical terminology, what is the sample size and what standard errors should we report? Standard errors are the way that statisticians indicate the amount of uncertainty in their averages, and they are the way we judge whether the average performance of one system is significantly different from the average performance of another one. Roughly speaking, differences in averages that are greater than twice their standard errors are deemed significant and those that are less than this are considered insignificant. The effective sample size is a crucial element in calculating the standard error.

The effective sample size indicates how many independent implementations we have of each kind of system, and the more independent implementations we have, the more likely it is that idiosyncratic factors are averaged out across the implementations.[38] The sample size depends upon the level at which the factors that affect voting system performance operate. If an idiosyncratic factor affects a whole county, then the number of counties must be considered the sample size. But if a unique factor operates independently on each individual, then the number of voters must be considered the sample size. For voting systems, idiosyncratic factors certainly operate at several levels including the individual person, the specific voting unit (i.e., a particular lever machine or Votomatic device), and the precinct. They may also operate at the

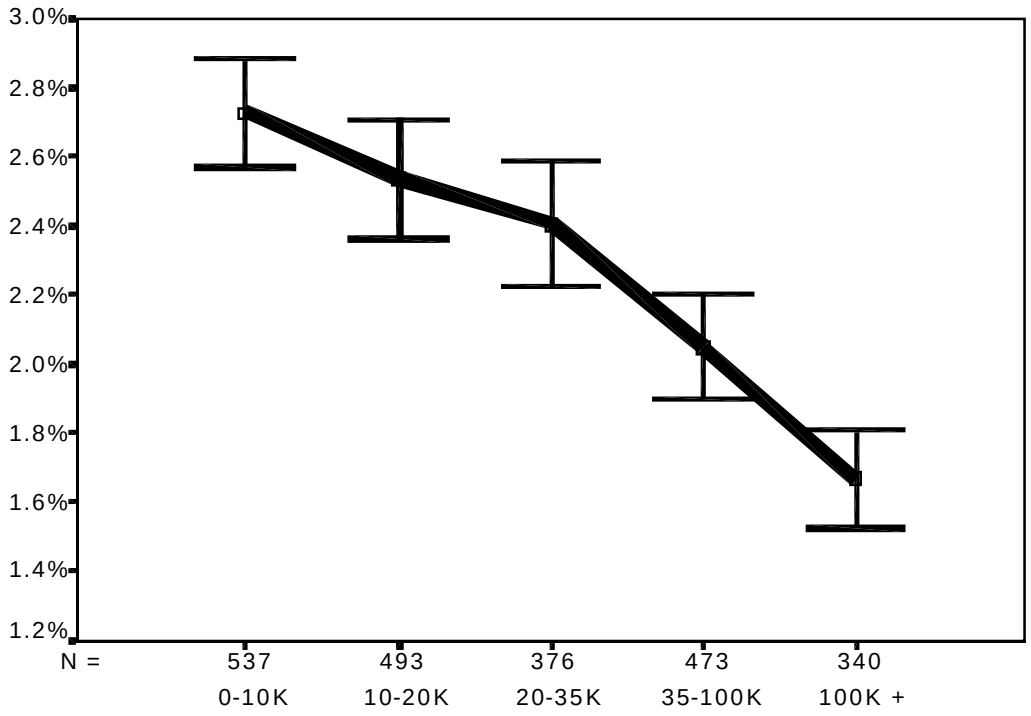
[36] Indeed, to complicate matters, the senatorial residual vote rate goes up with population size.

[37] The Calt-Tech/MIT technical report cites the number of counties as their sample size without any discussion or commentary.

[38] Technically, we can only be sure that averaging reduces the impact of idiosyncratic factors if systems have been randomly assigned to counties.

Figure 3

Presidential Residual Vote by County Population



Population of Counties in 1,000's of People

county and state level. Our dataset has eighty million voters, hundreds of thousands of voting devices, 135,000 precincts, 2,186 counties, and 38 states. How many units are we analyzing?

An extremely conservative approach suggests that we should assume that we are analyzing 2,186 implementations of voting technologies in the whole dataset and only 100 implementations when we consider the 100 most populous counties. But the major kinds of idiosyncratic factors—individual characteristics or machine malfunctions or even the vagaries of precinct administration—are undoubtedly averaged out to a greater extent in a large county than in a small one. Certainly it must mean more to find out that an electronic system has a presidential residual vote rate of 0.7% in a large urban county such as Clark County, Nevada with a population of three-quarters of a million, almost 400,000 voters, and 850 precincts than that the electronic systems in Lowndes County Alabama or Allen County Kentucky have the same residual vote rate in places with 13,000 people, 6,000 voters and a dozen precincts. This strongly suggests that we should put more emphasis upon larger counties when we consider the statistical reliability of results.

We will put more emphasis upon large counties in two ways. When doing our analysis of all the counties we will use what statisticians call "weighted least squares" methods that put more emphasis upon deviations from the average for large units than for small units. For the 100 most populous counties, we will use unweighted methods because, although they vary somewhat in size, they are much more similar in population than the entire set of counties in the United States.[39]

Weighted least squares methods ensure that we engage in a proper statistical accounting for the variations that occur from county to county. They put a smaller weight on variations in performance in small counties and a larger weight on variations in large ones because these variations have not been averaged out across large numbers of people, voting units, or precincts in smaller counties. This approach to dealing with statistical variability is based upon standard statistical practice which requires identifying where the variability is located. It is based upon our best judgment about where the variability is.

There is another statistical question, the issue of the representativeness of our results, where our preferences have greater sway because the answer depends upon our purpose. If we want our results to be representative of counties, then we average over counties. The average provides an answer to the question: "If I were to pick a county at random that uses a certain voting system, what is the best guess about its residual vote rate?" A naive voting administrator who knew nothing about his county might proceed in this way although we think, as we suggest above, that most administrators would ask about counties of the same size as theirs. That is why we present separate analyses of all counties (which are mostly small ones) and the 100 most populous counties.[40]

There is another issue that we might want to address as well. If we want our results to be representative of the voting population, then we weight by the voting population. This approach answers the following questions: "If I were to pick a group of voters at random who use a certain voting system, what is the best guess about their residual vote rate?" Or "What is the percent of all ballots cast using this system that were residual votes?" This question is the one that a voter or a policy-maker might ask. We provide an answer to this question by also presenting information about voting system performance that is weighted by the size of the counties.

Whichever questions we ask, it is still important to deal with statistical variation properly so that any inferences that are made from the data take into account the way that they create uncertainty in our results. Thus, whether we want our results to be representative of counties or voters, we take care to provide the best possible treatment of standard errors as discussed above.

---

[39] Except for Los Angeles California (population about eight million) and Cook County Illinois (population about five million), the range of populations in the 100 largest counties is from about 500,000 people to 2.5 million people—a factor of about five. Even including Los Angeles and Cook the ratio of largest to smallest is only about 16. Statistically speaking, these ratios cannot cause too much mischief. But among all counties, the smallest one on which we have data (Arthur County Nebraska) has a population of about 500 which is a factor of 1000 smaller than the smallest of the 100 largest counties, and ten percent of the counties have populations less than about 5,000 people which is a factor of 100 less than the smallest of the 100 largest counties and a factor of 1000 less than Cook County. Ratios of 100 to 1000 can cause problems. Note that approximately the same ratios affect voting units and precincts so that any idiosyncratic factors due to them will also be related to the population size.

[40] There is still an unresolved issue here that we skirt by making a relatively conservative assumption. We believe that it clearly makes sense to weight variation in performance among larger counties more heavily than about smaller counties, but we can think of a number of ways to think about the total sample size. We take the relatively conservative approach of assuming that it is equal to the number of counties, but a case could be made for using the number of states (because of different state laws and practices with respect to elections and election reporting), the number of precincts, or some other number. We use the number of counties because we suspect that there are unobserved county characteristics that have a bearing on the results that we get. The Cal-Tech/MIT method of looking at changes in counties over time helps to control for this, but our worries about data quality makes us very wary of their results.

Controlling for the Different Characteristics of Counties

All of the preceding considerations apply equally well to randomized experiments as to the observational data that we have. But because voting systems are not randomly assigned to counties, observational data add another set of concerns that we must consider carefully. There are many reasons to believe that the counties differ in important ways that affect residual vote rates. These factors can confound the analysis by making it seem as if voting systems differ when it is the people in the counties or some other factor that differs.[41]

Consider Figure 4 which shows the relationship between residual votes in the 2000 presidential election and high school graduation rates. The figure plots the average residual votes by seven categories each of which represent increases of 5% in graduation rates for the county. The "bars" around each mean value indicate a 95% confidence interval which is one way of representing uncertainty in the data. There is a clear downward trend with each 5% increase in graduation rates leading to roughly .3% decline in residual votes. This means that the residual vote rate will typically be significantly higher in those counties with lower rates of high school graduation. For the five types of systems, high school graduation rates are 74% for punchcards, 66% for DREs, 72% for optical, 65% for lever machines, and 73% for paper ballots. These differences can easily account for up to half a percentage point of the differences in residual vote rates.

Or consider Figure 5 which shows the relationship between residual votes in the 2000 presidential election and the total percentage of minorities including African Americans, Hispanics, Asian Americans, and Native Americans. As the percentage of minorities increases, the residual vote rate increases as well, especially in those counties with a large percentage of minorities.

Once we know that some county characteristics matter, we can try to correct for differences in them across counties by adjusting residual vote rates upwards or downwards. Statisticians have developed sophisticated ways for doing this such as multiple regression analysis. We will use this technique, but it is far from fail-safe. There are two problems in this context. One is simply figuring out all of the factors that might confound our analysis. The other is obtaining good measures of them.

In addition to voting system type, there are three classes of factors that might affect the residual vote performance of voting systems. These are the characteristics of voters, the characteristics of the county election administrations, and the characteristics of the political campaigns that draw people to the polls. Individual characteristics such as education, language, minority status, economic condition, newness to the United States, limitations on physical mobility, and age might have an impact on residual voting percentages (Walker 1966; Vanderleeuw & Engstrom 1987). These factors may influence people's capacity to understand how voting systems work and to operate them effectively. The characteristics of counties might also affect residual votes through the quality and character of election administration. Our imperfect measures for these factors include percent urban and total population. Finally, the closeness of political campaigns might also matter (Crain,, Leavens, & Abbot 1987). Unfortunately, it is very difficult to think of all the things that might confound our analysis.

[41] The literature on the causes of residual votes and spoiled ballots is relatively small and scattered, but it still suggests a substantial number of possible confounding factors. Residual votes have been explained in terms of political factors such as voter concern with the importance of the office (Dubois 1979), the voter's estimates of the relative closeness of races on the ballot (Crain, Leavens, & Abbot 1987), the voter's concern with one or another level of government (Cox & Munger 1989 p 225), voter mobilization efforts that bring less skilled or new voters into the voting booth, voter apathy and ignorance (Stiefbold 1965 p 406), voter demobilization due to negative advertising (Ansolabehere et al. 1994 p 833-34), or voter protest (Kim and Koh, 1972; Stiefbold 1965 p 406-7). "Roll-off" from people not voting for offices farther down the ticket has been attributed to the rise of industrial capitalism and its takeover of the American party system in 1896 leading to demobilization of the electorate (Burnham 1965 p 24-6) or as simply the result of the implementation of the Australian ballot (Rusk 1970 p 1237). Residual votes and roll-off have also been attributed to the position of candidate names (Bain & Hecock 1957; Darcy 1986; Kelley & McAllister 1984), to the partisanship of the ballot (Walker 1966 p 456-58, Dubois 1979 p 879-81), and to the type of voting machine (Thomas, 1968; Montgomery 1985; Darcy and Schneider 1989). Finally, residual votes and roll-off have been explained in terms of race, age, education, and income (Walker 1966 p 460; Vanderleeuw & Engstrom 1987 p 1091; Darcy and Schneider 1989 p 357-361) and their reduction have been attributed to increased voter education and sophistication (Mott 1926 p 881; Strong 1948 p 517; Barnes et al., 1962 p 906; Kim & Koh 1972, p 833).

Figure 4

Presidential Residual Vote by

High School Graduates in County

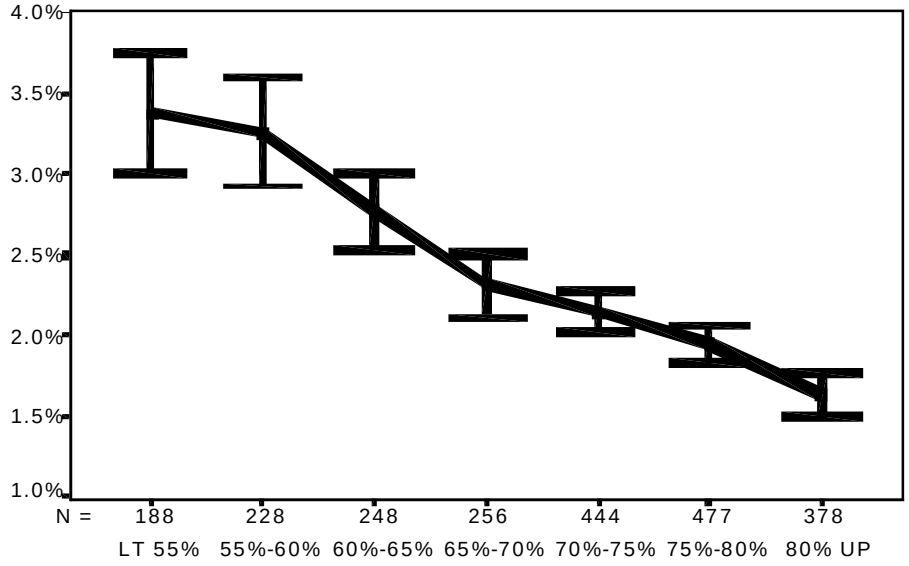
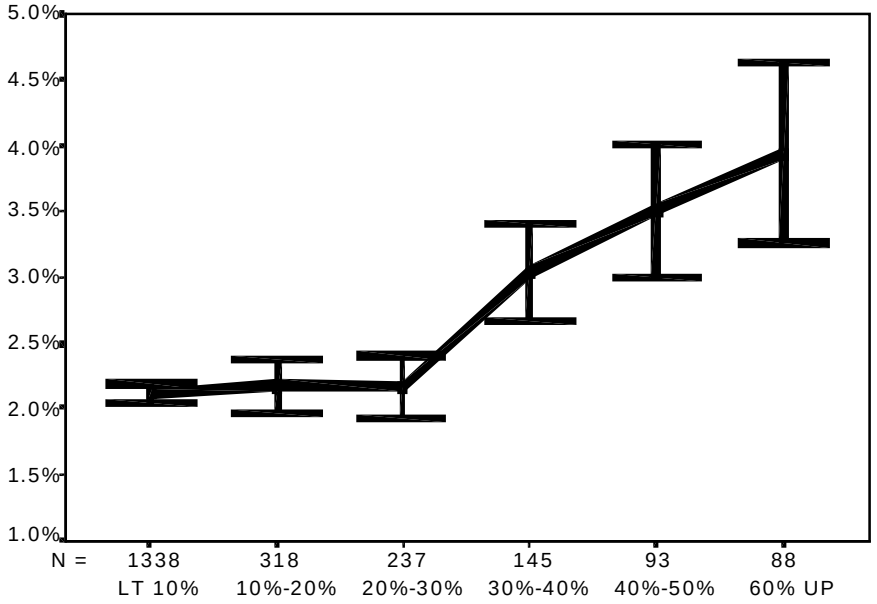


Figure 5

Presidential Residual Vote by Minorities in County



The second problem is that once we have thought of these things, we must obtain good measures of them. In doing this, we are typically constrained by the available data as described in an earlier section.[42]

There is no absolutely foolproof way to solve these problems, but statisticians have found that results are more likely to be true when they are robust against different assumptions and specifications. A finding is robust, and not fragile, when it is obtained using a variety of plausible assumptions, a variety of reasonable statistical methods, and a variety of different controls for confounding factors. Robust results have a greater likelihood of being true than fragile ones.

In future work, we intend to explore more sophisticated models for analyzing these data such as the over-dispersed binomial estimated using robust techniques used in a recent paper co-authored by one of the authors of this report (Wand et al, 2001). The techniques used in that paper take into account the different sizes of counties, the basic nature of residual vote data, the need to make adjustments for poor data quality, and the differences across states. Rather than using that very complicated method here, we will utilize simple averages, weighted averages, ordinary least squares with dummy variables to adjust for different state means, and weighted least squares. We believe that these methods, along with our many tests for robustness by using different subsamples of the data, provide a good basis for the rather modest conclusions that we make.

[42] Those knowledgeable about statistics might think that we should also discuss the ecological inference problem (Goodman 1953, 1959; Achen and Shiveley, 1995; King, 1997) that typically bedevils inferences from county or precinct level voting data. However, the ecological inference problem is related to the situation where we want to infer individual level behavior from county level data that includes mixtures of individuals with different characteristics. For example, we might want to know at what rate African Americans voted for a Democratic candidate when all we have is a sample of counties with different percentages of African Americans and different percentages of vote for the Democrat. Even if the percentage of Democratic vote appears to go up with the percentage of African Americans, it is possible that something other than blacks voting for Democrats could explain the observed data. In our case we have counties in which every voter (putting absentee voting aside) got the same voting system and we observe the percentage of residual votes. There is the possibility that contextual effects (such as some organization giving people in a county incorrect information about how to vote with some technology) might be a problem in making inferences in this situation, but there is no problem of ecological inference if residual votes are due to unintentional mistakes by voters. Even if some votes are due to intentional mistakes, unless there is a contextual effect from some organization telling people not to vote, there is, in principle, no greater difficulty in controlling for individual factors that cause intentional voting than with individual level data.

V. DESCRIPTIVE STATISTICAL ANALYSIS FOR PRESIDENTIAL RESIDUAL VOTES

Average Performance in all Counties

Table 1 presents residual votes by voting system type for the presidential election of 2000. The first column presents the average over all counties and the second the average over all ballots cast. The third column presents the standard deviation for the average over all ballots cast and the fourth indicates the number of counties.[43]

Table 1: Average Presidential Residual Votes by Voting System

Voting System	County Average	Average over all Ballots	Standard Deviation	Number of Counties
Punchcards	2.87%	2.64%	1.66%	435
DREs	2.36%	1.68%	1.19%	266
Optical Scan	2.13%	1.37%	1.48%	1018
Lever Machine	2.20%	1.72%	1.31%	323
Paper Ballot	2.23%	1.99%	1.42%	177
National Average	2.32%	1.94%	1.59%	2219

Only one system has a remarkably different level of residual votes—punchcard systems. Every other system is at least half a percentage point better no matter which measure is used, and if the average over all ballots is considered, every other system is from two-thirds to a full percentage point better. For the remaining systems, their relative order depends upon which measure is used, and the absolute differences are on the order of one-fifth to one-third of a percentage point. When the county average is used, the best systems are optical scanners and they are about one-fourth of a percentage better than DREs which have the highest rates of residual ballots after punchcards. When the average over ballots is used, paper ballots have the highest rates of residual ballots after punchcards and DREs are the second best system next to optical scanners. The differences among the non-punchcard systems appear to be small given our concerns about data quality and the many factors that could confound these results. Although statistical tests might suggest that these differences are statistically significant,[44] the assumptions required for those tests are simply not met with these data, and we do not consider one-fifth to one-third of a percentage point to be worthy of note. Earlier we showed that the use of different systems for absentee and election-day voters could change the residual vote rate by several tenths of a percentage point. Differences across counties and states in reporting protocols regarding total ballots could have the same kind of impact. Consequently, differences of one-fifth to one-third of a percentage point could easily be attributed simply to data quality. It is only when we see persistent differences on the order of half a percentage point that we are willing to take notice. Only the comparison of punchcards with other systems provides this kind of evidence for significant and substantial differences.

Even after taking into account data quality, the problems of confounding must be considered because of variations across counties in the education level of the electorate, the quality of election administration, and the perceived closeness of the presidential election. These problems are addressed in later sections of this report.

These data do not at all support the contention of the Cal-Tech/MIT report that DREs lost nearly as many

[43] As discussed in an earlier section, the standard deviation is computed using a weighted average of all voters.

[44] One measure of the standard error for each of the two averages in Table 1 is the standard deviation divided by the square root of the number of counties. Thus, the standard error is .08% for punchcards, .07% for DREs, .05% for optical scan, .07% for lever machines, and .11% for paper ballots. From this calculation, differences of .10% to .20% and larger might be considered significant, but using these figures requires the assumptions that the number of counties is the right sample size and that there are no biases due to the data or to confounding factors.

votes as punchcards and that paper ballots have performed better. We are, of course, using only 2000 data while the Cal-Tech/MIT report uses information from the 1988, 1992, and 1996 presidential elections. One explanation is that DREs have gotten better. Another is that the data from those earlier years is less trustworthy than the 2000 data, as noted earlier.

Poorly Performing Systems

Averages are very interesting, but it is also important to know something about which systems sometimes perform very badly. The standard deviations in Table 1 provide some information about this because they tell us how much performance varies for these systems. Most of this variance has to be in high residual vote rates because the lowest possible rate is zero. The lowest standard deviation is for DREs. The reason is that a very small percentage of electronic systems perform very poorly based upon the Cal-Tech MIT standard of having more than 5% residual vote. Only 4.5% of counties with DREs perform this poorly—the corresponding figures for other systems are 9.9% of counties with lever machines, 9.0% of those with punchcards, 7.5% of those with optical scan systems, and only 4.0% of those with paper ballots. These figures are somewhat surprising considering that the Cal-Tech/MIT report singled out DREs for having very high residual vote rates.[45]

Indeed, in our search for poorly performing systems, we found that thirty-five counties reported residual vote rates over 8% and none of them used DREs, but a disproportionately high 66% of the counties used optical scans (46% of all counties use optical scan systems)[46] and a somewhat high 23% used punchcards (20% of the population uses punchcards). The remaining 11% used lever or paper systems. One explanation for this, of course, is that these are all small counties in which we would expect some extreme values, but we also found nine counties with a population over 100,000 that reported residual vote rates over 5% and none of these used DREs—eight of them used punchcards (six Votomatic, one Datavote, and one Vote Recorder) and one used Accuvote's precinct count optical system.

Table 2: Average Presidential Residual Votes by Voting System
(Rows Sum to 100% Except for Rounding)

Voting System	Good	Adequate	Worrying	Unacceptable	Number of Counties
Punchcard	6.4%	22.3%	35.6%	35.6%	435
DREs	20.7%	22.6%	25.6%	31.2%	266
Optical Scan	32.8%	28.5%	16.4%	22.3%	1018
Lever Machine	31.0%	29.7%	13.6%	25.7%	323
Paper Ballot	17.5%	36.2%	24.3%	22.0%	177
% With Rating	24.7%	27.4%	21.5%	26.5%	2219

[45] "We also believe that optical scanning dominates older full-faced, push-button DREs, which comprise fully two-thirds of the electronic machines in our analysis. Touch screens are, in our opinion, still unproven. Some counties, like Riverside, California, have had good experiences; other counties like Beaver County, Pennyslvannia, and many counties in New Mexico had very high residual vote rates (over five percent in 2000)." (Cal-Tech/MIT, 2001, p 22). These sentences seem to imply that touch screens performed badly in New Mexico, but in fact, as far as we are aware, all of the electronic systems in New Mexico were push-button DREs.

[46] Almost half (11 of 23) of these optical scan systems were precinct scanning systems.

Good, Adequate, Worrying, and Unacceptable Performance in All and in Large Counties

Average performance, clearly, is not the whole story. Table 2 uses the National Commission on Federal Election Reform's four point evaluation system to compare the five types of systems for all counties. The table is constructed so that the entries are row percentages which indicate the percentage of each type that is good, adequate, worrying, or unacceptable. If a system's performance in many counties is good, then the row percentage in the "good" column will be high; if it is inadequate in many counties, then the entry in the "inadequate" column will be high.

The story for all U.S. counties is similar to the one reported in Table 1. Punchcards perform poorly with only 6.4% of the counties in the "good" category and only 29% either good or adequate. Over a third of the punchcard counties have unacceptable performance. All the other systems perform appreciably better. About 43% of the counties using DREs reach the adequate or good standard, about 54% of those counties using paper ballots reach this standard, and 61% of the optical scan and lever machines are adequate or good.

These data are very useful, but they will be misleading if some systems work well in small counties but not in large ones. Paper ballots, for example, work relatively well, but they are implemented in small counties serving less than one percent of the U.S. population. Yet about 40% of U.S. voters are in the 100 counties with populations larger than 500,000 people.[47] Both the Carter/Ford Commission and the Cal-Tech/MIT report singled out the 40 largest counties for special attention,[48] and it makes sense to consider large counties separately.

Table 3 shows how voting systems perform according to the Carter/Ford Commission standards for 77 of the 100 largest counties on which we have all the requisite information.[49] The results show that among the large counties, about three-quarters of the DRE and optical scan counties are rated good which is much better than the other two systems, punchcards and lever machines, that are used in large counties. Indeed, the simple (unweighted) residual vote average is .79% for optical scan systems and .96% for DREs. Levers are at 1.47% and punchcards are at 2.48%.

Table 3: Average Presidential Residual Votes by Voting System for Largest Counties

(Rows Sum to 100% Except for Rounding)

Voting System	Good	Adequate	Worrying	Unacceptable	Number of Counties
Punchcard	17.2%	27.6%	31.0%	24.1%	29
DREs	72.7%	18.2%	9.1%	0.0%	11
Optical Scan	77.8%	22.2%	0.0%	0.0%	18
Lever Machine	36.8%	31.6%	5.3%	26.3%	19
Paper Ballot					0
% With Rating	44.2%	26.0%	14.3%	15.6%	77

[47] According to the 1990 Census data the 100th largest county in the United States, Ramsey County, Minnesota where St. Paul is located, had a population of 486,000 people. All but three of the 100 largest counties had populations in 1990 of over 500,000 people. The 100 largest counties comprised 43% of the U.S. population, and somewhat less of the voting population.

[48] The National Commission on Federal Election Reform (2001) classifies the 40 largest counties according to its rating system on page 55. The data come from the Cal-Tech/MIT report (2001) which provides a table of residual votes for the 40 largest counties on page 90. We added more counties to make sure that the pattern of performance across voting system types for those 40 counties remained the same, which it does, as more counties are considered.

[49] Thirteen of these 100 have a mix of voting systems. (Six in Massachusetts, four in Michigan, and three in Texas.) We do not have total ballots for the remaining ten. (One in Indiana, two in Missouri, two in Oklahoma, two in Pennsylvania, two in Tennessee, and one in Wisconsin.)

One way to interpret these results is that optical scan and DREs in large counties deserve the highest rating of "good" based upon the scale put forth by the Carter/Ford Commission.[50] Although the average for residual votes for lever machines might be characterized as "adequate," the large number of counties in the "unacceptable" category suggest caution in making this statement. Punchcards are "worrying" based upon their average residual vote of 2.48% and the fact that 55.1% of them are in the last two categories.

What can we conclude from these results so far? We must emphasize the need for caution in order to avoid coming to hasty conclusions, and the rest of this report provides additional analysis that elucidates the simple results provided by the descriptive statistics reported here. But, after perusing the data as carefully as we have done, we owe our readers a summary to this point.

The consistent thread through this discussion is that punchcards always perform the worst and the other four systems always do much better. We feel very confident about this conclusion based upon our entire analysis. In this section and in the remainder of the report, we always find that Votomatic style punchcards[51] perform poorly and nothing seems to change that conclusion. Based upon our analysis we believe that a move away from Votomatic style punchcards will, barring very unusual circumstances, typically result in a substantial reduction in residual votes.

The rest of our conclusions are more tentative and should be taken as working hypotheses about the relative merits of these systems. We are confident that the statements in the next paragraph correctly summarize the pattern of residual votes in 2000 across the different kinds of systems, but we are less sure about whether or not that pattern is the result of the inherent characteristics of the systems or a result of the circumstances under which they have been implemented. Thus, the poor performance in some counties may have much more to do with the conditions in those counties than with the relative merits of the systems. Consequently, we would not want to recommend that a county move from one non-punchcard system to another based upon our findings. Instead, we want to suggest that our findings are working hypotheses about the relative strengths and weaknesses of these systems that deserve further scrutiny using more data and better research designs.

With these caveats, we can say the following. Paper ballots typically do well, but they are only used in small counties (the largest has a population less than 60,000 and the average population is 6100 people) and the data tell us virtually nothing about what would happen if they were used in larger counties. Lever systems are competitive with optical scan systems and DREs across all counties but they seem prone to very poor performance and they do not do well in large counties. Optical scan and DREs appear to dominate all other systems in terms of overall performance across all counties and especially in large counties. DREs do not do as well as optical scan systems in smaller counties, but optical scan systems seem slightly more prone to very poor performance than DREs. The data suggest that we need to know much more about all of these systems because about a third of the population and more than half the counties in America have systems that are "worrying" or "unacceptable."

In the remainder of this report we explore other measures of performance and we describe a multivariate statistical analysis that controls for a number of possible confounding factors. We also discuss two California case studies that show the improvement obtained by introducing optical scan and DRE systems. We end with findings and recommendations.

---

[50] We base this statement upon the fact that about three-quarters of all counties using each system fall into this category and the overall average for these counties is less than 1% residual votes.

[51] This qualification is meant to indicate our uncertainty about the merits of Datavote style punchcards. This issue is discussed in more detail later in the report.

VI. MULTIVARIATE ANALYSIS FOR PRESIDENTIAL RESIDUAL VOTES

We use multivariate statistical analysis to make corrections for confounding factors because voting systems have not been assigned at random to election jurisdictions in the United States. Consequently, the observed differences in the average performance of each voting system could be the result of differences in voting jurisdictions instead of the characteristics of the systems themselves. In this situation, researchers typically try to use statistical methods, such as multiple regression analysis, to control for those characteristics that might confound the analysis. We have already identified two such characteristics, percent high school graduates and percent minorities in a district. There is an endless list of other possibilities which means that no one can ever claim to have controlled for all factors that might confound the results of a study like this one. At best, researchers can try to think of everything that might matter and try to collect data on these things. This effort is inevitably limited by the available data.

Correcting for Confounding Variables

To correct for confounding factors we have used census data on the characteristics of people and counties and political data on the competitiveness of races. We have considered a large number of census variables including the following: percent high school graduates, median family income, percent over age 65, percent with mobility limitations, percent who use a language other than English, percent owner occupied housing, percent in poverty, percent female headed households, percent entering the county between censuses, percent urban, population of the county, etc. We have also considered a number of political variables described earlier. And in some specifications we have included separate dummy variables for each state to account for baseline differences from one state to another.

In the end, we have settled upon a rather small set of variables that are consistently important and that capture most of the variance in residual votes. We do not claim that our final specification using these variables corrects for all possible confounding influences, although we do know that our results for presidential residual votes do not change very much by adding the other variables listed above. Nevertheless, there may be some other important factors that we have not included in the specification, and there is the possibility that the factors we have included operate differently in different counties. Consequently, we strongly believe that our results are not the last word.

We have included three socio-economic factors in our multivariate equations: the fraction of the county's population that is high school graduates, the fraction of minorities, and the median family income (in $100,000 units so that the average income of families in the counties runs from .11 to .65). We have included three political variables: competitiveness of the presidential race in the state, competitiveness of the Senate race (if there was one), and competitiveness of the governor's race (if there was one). For the presidential, senatorial, and gubernatorial races, the measures were coded from zero (uncompetitive) to one (highly competitive) with a coding of zero for the senatorial measure if there was no Senate race and a coding of zero for the gubernatorial measure if there was no Governor's race.

We expected that greater competitiveness in the presidential race within a state would decrease *presidential* residual votes, although we did not expect a very large effect because the presidential race is essentially a national one, and we expected that greater competitiveness in the senatorial race or gubernatorial race would somewhat increase *presidential* residual votes because some voters would come out to vote in the senatorial or gubernatorial race but not the presidential race. We expected competitiveness in the senatorial race to decrease senatorial residual votes because people who might otherwise come out just to vote for president would be drawn into the senatorial contest, and that competitiveness in the presidential race would increase senatorial residual votes because some people would come out just to vote in the presidential race. (There were only seven states where there were both senatorial and gubernatorial races, so we did not expect that we would be able to detect much impact of the competitiveness of gubernatorial races on senate voting.)

[52] The states are Delaware, Indiana, Montana, North Dakota, Utah, West Virginia, and Washington.

Finally, we included an indicator variable for each type of voting system that is scored as one if the county uses that system and zero otherwise. Following standard practice, we chose one system, in our case optical scanners, as our statistical baseline system so that all reports below are the deviation in the performance of the given system from optical scan systems. We use optical scan as the baseline because positive coefficients for the other systems will indicate that they produce more residual votes than optical scan systems and negative coefficients will indicate that they produce fewer residual votes.

For the analysis of all the counties, we used weighted least squares where the weight was the total number of ballots and the estimation routine was allowed to choose the best power value for the weight. For the analysis of the large counties, we used standard ordinary least squares.

Presidential Residual Votes in All and in Large Counties

Table 4 presents our results for presidential residual votes for all counties and for large counties. We have converted the coefficients for the systems into percentages which indicate the amount by which each system exceeds the statistical baseline system (optical scanners). The quantities in parentheses are standard errors. Thus, punchcards have .98% (essentially one full percentage point) more residual votes than optical scan systems and the standard error is .09% or slightly less than one-tenth of a percentage point.

Table 4: Presidential Residual Votes with Controls for Confounding Factors
Regression Coefficients and Standard Errors (in Parentheses)

Factors in the Analysis	Analysis for All Counties		Analysis for Large Counties	
SYSTEM TYPE				
Optical Scan	**Baseline System**		**Baseline System**	
Punchcard	.98%	(.09%)***	1.93%	(.40%)***
DREs	.09%	(.11%)	0.08%	(.54%)
Lever Machine	-.38%	(.11%)***	0.73%	(.53%)
Paper Ballot	.12%	(.16%)	No paper systems	
COUNTY FACTORS				
Fraction H.S.Graduates	-.039	(.005)***	-.020	(.040)
Fraction Minorities	.020	(.002)***	.024	(.012)*
Median Family Income	-.034	(.006)***	-.027	(.027)
POLITICAL FACTORS				
Pres. Competitiveness	.000	(.001)	.005	(.005)
Sen. Competitiveness	.004	(.001)***	.017	(.005)*
Gov. Competitivenss	.003	(.001)**	.011	(.016)
MODEL STATISTICS				
R^2	.194		.448	
Number of Counties	2219		77	

Note: Asterisks indicate significance levels. * (.05), ** (.01), *** (.001)

Although the results for the voting systems are of primary interest, it is worth noting that all of the control variables have the expected signs and magnitudes. As the proportion of high school graduates or the median family income goes up in a county, residual votes go down; an increase in the proportion of minorities leads to an increase in residual votes. Each of these variables is coded so that its lowest value is zero and its highest value is no more than one. For the demographic variables, a change of .10 is a significant one because it indicates an increase of 10% in the proportion of high school graduates or minorities and an increase of $10,000 in average income. For the political variables, a change of .30 is a significant increase in competitiveness.

Across all counties, we find that an increase of .10 in the proportion who graduated from high school leads to a .39% decrease in residual votes. An increase of .10 in the proportion of minorities leads to a .20% increase in residual votes, and an increase in average income of $10,000 leads to a reduction in residual votes of .34%. All these figures are highly statistically significant for the 2219 counties. The coefficients for the demographic variables are similar for the large counties, but given the small sample size, only one of them (for fraction minorities) is statistically significant.

The impacts of political competitiveness are mixed. Presidential residual vote goes up with senatorial competitiveness and with gubernatorial competitiveness as we expected, but it is not affected by the competitiveness of the presidential race. The senatorial figure suggests that a .30 unit increase in senatorial competitiveness leads to a .12% increase inpresidential residual vote in all counties and a .51% increase in residual vote in large counties. These substantial results demonstrate the importance of controlling for these factors.

Except for the same recurring story about high residual votes for punchcards, the coefficients for the voting systems for all counties suggest a much different perspective than Table 1. In Table 1 the best to worst systems were optical scan, lever machines, paper, DREs, and punchcards. Here, the systems, from best to worst are lever machines, the trio of optical scan, DREs, and paper ballots which are not significantly different from one another, and punchcards. Lever machines appear to be significantly better than optical scan systems, and punchcards are quite a bit worse. The size of the differences among optical scan, DREs, and paper is about .1% which is statistically insignificant and substantively negligible. When we move to large counties, lever machines no longer look so good, but statistically there is no difference among DREs, lever machines, and optical scan systems. Punchcards look even worse than they do for all counties with about a 2% higher residual vote rate than the two best systems, optical scan and DREs.

These results are typical. We have never encountered a specification in which punchcards look good, and we have not encountered any specifications in which optical scan, DREs, or lever machines look much different from one another, although there is some indication that lever machines do not perform as well as these other two systems in large counties.

VII. SENATORIAL RESIDUAL VOTES

The Cal-Tech/MIT report also analyzes senatorial and gubernatorial residual votes. We have already expressed our uneasiness with using these residual votes as indicators of the performance of voting systems, and we present an analysis of senatorial residual votes here primarily as a way of illustrating the difficulties. The major complication is that although *some* senatorial and gubernatorial residual votes are undoubtedly unintentional undervotes or overvotes caused by voters' experiencing problems with voting systems, others are simply the result of voters deciding not to vote in the race. As a result, a fair comparison of average residual votes across systems must find a way to determine how much of the senatorial residual vote is unintentional and how much is intentional.

There is certainly strong evidence that part of the senatorial residual vote is intentional. If we dichotomize races into competitive and noncompetitive, we find that the senatorial residual vote is 3.0% for competitive races and 5.2% for noncompetitive ones.[53] This suggests that at least two percentage points of the senatorial residual vote in noncompetitive races is the result of intentional undervoting. (There is probably not much intentional overvoting.) It also suggests that we can use the competitiveness ratings to make some adjustments for this undervoting. But we know so little about intentional undervoting that we cannot specify the exact functional relationship between competitiveness and residual votes, and we do not know how competitiveness interacts with other characteristics of the voting public. For example, is there likely to be more intentional undervoting among those with lower education or with other characteristics? How should we specify this relationship?

To complicate matters, a complete analysis must not only adjust for differences in intentional undervoting from one county to another, it must also adjust for county characteristics that produce differences in unintentional residual votes. To make both adjustments correctly at the same time is asking quite a bit, and until we know more about how to do this, we must remain skeptical about the use of senatorial residual votes as a measure of performance. Nevertheless, for completeness and as a way to demonstrate the difficulties, we present an analysis of them.[54]

Table 5: Average Senatorial Residual Votes by Voting System

Voting System	County Average	Average over all Ballots	Standard Deviation	Number of Counties
Punchcards	5.29%	5.30%	3.16%	283
DREs	4.61%	4.61%	1.90%	78
Optical Scan	3.29%	3.50%	2.64%	610
Lever Machine	5.85%	4.93%	3.95%	196
Paper Ballot	3.61%	3.67%	3.33%	113
National Average	4.23%	4.58%	3.18%	1280

Table 5 presents average senatorial residual votes by voting system. A comparison of the national average for senatorial residual votes in this table with the corresponding figure for presidential residual votes in Table 1 indicates that senatorial residual votes are about 2% to 2.5% higher depending upon the average

[53] We simply assumed that those states with competitive rankings on our index below .50 were noncompetitive and those at .50 or above were competitive.

[54] We do not present an analysis of gubernatorial residual votes because there many fewer gubernatorial races than senatorial ones and the results are not that much different from those for senatorial residual votes.

that is used. Table 5 also shows that once again, punchcards do quite badly with the worst average (5.3%) over all ballots and the second worst average (5.29%) over all counties. Lever machines exchange places with punchcards as the worst (5.85%) or second worst (4.93%) system. DREs are not as bad as punchcards and lever machines, but they also have high rates of senatorial residual votes of 4.61%. Optical scan and paper ballots do the best with residual vote rates between 3.29% and 3.67%.

The averages in Table 5 are interesting, but we know that the performance of voting systems can change when controls are introduced for confounding factors. Table 6 shows what happens when statistical controls are introduced as before.

The results in Table 6 are not as clear as those we obtained for presidential residual votes, and we have found that the coefficients in Table 6 are very sensitive to the particular specification. Consider first the regression with all counties included in the analysis. Two of three demographic variables are statistically significant with the signs and sizes we would expect. The measures for presidential and senatorial competitiveness not only have the right sign (more senatorial residual votes when the presidential race is competitive and fewer when the senatorial race is competitive), but they are also quite sizeable. In fact, they are probably too big to be believable because a .3 increase in competitiveness in the presidential race (on the zero to one scale of competitiveness) increases senatorial residual votes by about eight percent and the same increase in competitiveness in the senatorial race decreases senatorial residual votes by about twelve percentage points. These numbers seem too big, and they suggest that the specification is having some problems capturing what is going on. The problem almost certainly stems from the difficulty of finding the right form and specification for adjusting senatorial residual votes for intentional undervoting. When results change dramatically from specification to specification as they do in this case, it is a sure sign that they should not be given a lot of credence.

Nevertheless, it is worth noting that the coefficients for the voting systems suggest that optical scan systems are the best, DREs the next best, then paper, punchcards, and finally lever systems. This is the first time that punchcards have done appreciably better than another system, but it only because lever systems look positively terrible, even worse than in Table 5 which presented the simple averages for each system. The other important change from Table 5 is that DREs have the lowest rate of senatorial residual votes next to optical scan systems after controlling for confounding factors. The major reasons why DREs improve is that counties using them have a low level of high school graduates (66% compared to a national average of 70%), and they are in states with very uncompetitive senate races (a value of .25 on our index compared to a national average of .44). When adjustments are made for these factors, DREs improve significantly. But we do not want to make too much of these results.

Even more than with the presidential residual vote analysis, we want to be very cautious about claiming very much for the results in this section. For the regression analysis to work, the control variables must adjust the residual votes in each county to take into account the competitiveness of the senatorial and presidential elections which affect *intentional* undervoting, and they must adjust for the different characteristics of the counties which affect *unintentional* undervoting. We simply do not know enough to make both adjustments simultaneously. Our lack of knowledge is reflected in the fragility of the specifications. For example, alternative specifications have led to results where DREs are insignificantly different from optical scan systems with a difference in residual vote rates of only about .5%. We do find that lever machines consistently perform very poorly and punchcard systems never do any better than their poor performance in Table 6, but the specification is definitely not robust.

One way to summarize this exercise is to say that there is some evidence that lever and punchcard systems perform poorly in recording senatorial votes and that other systems perform better, but there is no consistent finding about the relative performance of optical scan, DRE, and paper ballots. Another way to summarize it is to say that we should not use senatorial or gubernatorial residual votes as a standard for evaluating these systems until we know a lot more about intentional undervoting.

Table 6: Senatorial Residual Votes with Controls for Confounding Factors
Regression Coefficients and Standard Errors (in Parentheses)

Factors in the Analysis	Analysis for All Counties	Analysis for Large Counties
VOTING SYSTEM		
Optical Scan	Baseline System	Baseline System
Punchcard	1.93% (.20%)***	0.86% (1.01%)
DREs	1.03% (.27%)**	0.47% (1.39%)
Lever Machine	4.16% (.27%)***	2.18% (1.30%)
Paper Ballot	1.77% (.34%)***	No paper systems
COUNTY FACTORS		
Fraction H.S.Graduates	-.063 (.013)***	-.001% (.001%)
Fraction Minorities	.032 (.006)***	-.007 (.030)
Median Family Income	.003 (.014)	.004 (.068)
POLITICAL FACTORS		
Pres. Competitiveness	.026 (.002)***	.011 (.013)
Sen. Competitiveness	-.039 (.002)***	-.042 (.012)***
Gov. Competitivenss	.002 (.003)	-.028 (.062)
MODEL STATISTICS		
R^2	.364	.257
Number of Counties	1280	68

Note: Asterisks indicate significance levels. * (.05), ** (.01), *** (.001)

VIII. PRESIDENTIAL RESIDUAL VOTES AND VOTER EDUCATION LEVELS

The Cal-Tech/MIT report states that:

"We believe that the high rate of residual votes of DREs stems from the user interfaces. We have examined many of these machines. The mechanics of voting on these machines are often confusing (p 23)."

Our report, of course, does not find the same high rate of residual votes for DREs, but it is worth taking seriously the idea that some systems might be harder to use than others. If this is so, then we would expect that voters with lower education would have the most difficulties.

Consequently, we would expect that those counties with fewer high school graduates would have more residual votes than those counties with more high school graduates. Figure 4 shows that this is generally true.

But if one system is more difficult to use than another, then we should also find that there is an interaction between that type of system and those with lower education that leads to more residual votes. The slope of the line in Figure 4 should be greater for difficult to use systems than for easy to use systems. Statistically we can uncover this interaction effect by constructing an "interaction term" consisting of the product of the proportion of high school graduates and an indicator for that system. In a multiple regression, this term should have a negative sign because residual votes will go down as the counties have more high school graduates who can cope with the difficult system.

Table 7 presents a regression with these interaction terms (listed at the bottom of the table) for all systems. This regression also includes terms for each system for all but the baseline system. These terms are listed at the top of the table. (Note that the coefficients for the systems are not in percentages here so they are about 100 times smaller than those in the previous tables.) The coefficients of these terms indicate what the relative rankings of the systems would be in a county with zero percent high school graduates.[55] Because there are no such counties (the lowest proportion of high school graduates is about 32%), it is dangerous to extrapolate to such a mythical place.

As expected, the coefficient for the interaction of DREs with the proportion of high school graduates has a negative sign. But, except for paper systems (which are surely different than the rest), all the other systems have a much more negative coefficient. The coefficient for punchcards, in fact, is six times bigger than that for DREs. Furthermore, the coefficient for the interaction of DREs with the proportion of high school graduates is statistically insignificant whereas the coefficients for all the other interaction terms (again, except for paper) are highly significant and much larger in absolute value. These negative coefficients mean that the presidential residual vote gets smaller and smaller as the proportion of high school graduates increases in a county.

The overall operation of each system must be characterized both by its performance in the mythical county with no high school graduates and by its performance as more and more high school graduates enter the county. Ideally we want a system with a low residual vote rate for those counties with few high school graduates, and we want the system to maintain that performance as it is moved to counties with more high school graduates. We might expect a slight decrease in residual votes as we moved to these counties with more high school graduates (which would amount to a slight negative slope in a regression equation), but we would hope that residual votes were not much affected by the fraction of high school graduates in a county.[56]

[56] One possibility, of course, is that intentional undervotes decrease with education so that the interaction term picks up that effect as well as unintentional undervotes. Thus some slope in the line may be inevitable and it should not count against voting systems. But differences in these slopes can only be the result of different voting systems and this should count against those systems which produce an increased slope. There is no reason why voting systems should contribute to unintentional undervotes just because some people do not have a high school education.

Table 7: Presidential Residual Votes and the Interaction of System Type with Education
Regression Coefficients and Standard Errors (in Parentheses)

Factor Affecting Residual Vote	All Counties	
VOTING SYSTEM		
Optical Scan	Baseline System	
Punchcard	.042	(.008)***
DREs	-.015	(.007)*
Lever Machine	-.018	(.007)
Paper Ballot	-.038	(.013)**
COUNTY FACTORS		
Fraction Minorities	.019	(.003)***
Median Family Income	-.034	(.007)***
POLITICAL FACTORS		
Pres. Competitiveness	.000	(.001)
Sen. Competitiveness	.003	(.001)**
Gov. Competitivenss	.002	(.001)
SYSTEM TYPE INTERACTING WITH EDUCATION		
Optical Scan X High School Grad.	-.000375	(.000061)***
Punchcard X High School Grad.	-.000805	(.000105)***
DREs X High School Grad.	-.000138	(.000096)
Lever Machines X High School Grad.	-.000409	(.000102)***
Paper Ballots X High School Grad.	.000157	(.000167)
MODEL STATISTICS		
R^2/Number of Counties	.186/2219	

Note: Asterisks indicate significance levels. * (.05), ** (.01), *** (.001)

The information in Table 7 suggests that no system performs as well as we would like, but some systems perform worse than others. Once again, punchcards are very worrisome—in fact, inadequate by the National Commission's definition. Consider a county that is average on all of the factors in Table 7 except that it has 56% high school graduates. This proportion of high school graduates would put the county in the bottom decile of counties with respect to high school graduates. Assume further that the county uses a punchcard system. Then the equation in Table 7 predicts that there will be 4.5% residual votes in the county compared to 2.7% that would result from an optical scan system and 2.5% for an electronic system. In a similar county with 63% high school graduates (the bottom quartile of counties), the residual vote rate for punchcards would be 3.9% and the residual vote rate for DREs and optical scan systems would be roughly equal at 2.4%. Even in a county with 90% high school graduates, punchcards would still be dead last with 2.3% residual votes compared to 2.1% for DREs and 1.7% for optical scan systems.

These numbers should be treated with caution, but they certainly suggest two things: DREs are not

especially hard to use compared to other systems because there is little interaction between residual votes and education for them, and punchcard systems appear to be very hard for everyone to use—even those with higher levels of education have trouble with them.

These results seem to contradict the Cal-Tech/MIT claims about DREs. Of course, the authors of the Cal-Tech/MIT report say that they base their claim upon their examination of these systems, but we have also examined, and in some cases voted with, DRE systems, Votomatics, optical scan systems, and lever machines. Our impression is that all of them need improvement, but we do not believe that DREs can be called the most difficult or even more difficult than the others.

In the end, we believe that careful experiments in which people are randomly assigned to system type are the only way to resolve this question. The study of Susan King Roth (1998) is a good beginning. In her comparison of lever and electronic machines, she concluded that:

"Results of the study were supplied to election officials with the suggestion that further studies be conducted on other systems. Soon afterwards electronic machines were purchased for use in the county [Franklin County, Ohio], replacing mechanical lever machines."[57] (Roth, 1998, 5)

Roth's study concluded with the call for more studies of the "usability and accuracy of voting systems." (1998, 8). We agree.

[57] In our data, Franklin County Ohio had a presidential residual vote rate of .89% and a Senate residual vote rate of 6.0% in a state with a noncompetitive Senate race and with a mean senate residual rate of 5.8%. (Ohio has 70 counties using punchcards, 5 using DREs, 11 using optical scan, and 2 using lever machines.)

IX. FINDINGS ABOUT SPECIFIC SYSTEMS

Although we have noted that there are different kinds of punchcard, optical scan, and electronic systems, our analysis has used five broad categories to classify voting systems. We have been forced to do this by the quality of the data. In this section, we bring to bear a number of different data sources to make some points about these different systems.

Twenty-one California counties use the Datavote system and eight (including Los Angeles) use some variant of the Votomatic system. In the 2000 presidential election in California, the residual vote rate for the Datavote counties was .85%, and it was 2.23% for the Votomatic style counties. These data suggest that Datavote can do quite well under some circumstances and that its performance might be significantly different from the Votomatic systems. California data also suggest that precinct optical scan systems do better than central count optical scan systems.

We have tried to check these findings against the national data, and we do find hints of these patterns, but the results are not persuasive. The problem is the national data simply cannot support inferences that go much beyond the conclusions that have already been described because, among other things, the information on subtypes of systems is not complete.

In addition to using standard ways to classify voting systems, we also explored whether the data suggested some new ways to classify and understand voting systems. One indication that there might be a useful way to break down a classification is when the members of the class form two clusters along some dimension. An examination of presidential residual votes for DREs suggest that some DREs cluster in a better performing group than the others. After trying various ways of breaking DREs into two groups that captured this difference, we found that the best distinction seemed to be between DREs in states where straight party line voting is allowed versus those in states where it is not allowed. The 221 DREs in party line states have an average residual vote percentage of 2.63% which is two and one-half times bigger than the average residual vote percentage of 1.02% for those DREs in non-party line states.

We do not have a good explanation for why DREs cluster in this way, but we have a hypothesis that bears further examination. DREs such as Microvote can be configured in two different ways to handle straight party line voting. The configuration option is selected when the system is purchased, and it can only be changed by a qualified repair person. One option is that a party line vote includes votes for all offices ("the one button option") and the other option is that a party-line vote includes votes for all offices *except* the president, and presidential votes must still be cast separately. North Carolina and South Carolina mandate by state law that DRE systems must be set for the latter option. As a result, these states go to substantial efforts to educate voters about this option so that voters who want to vote straight party tickets are sure to notice. After all, if voters are not clearly notified, they may mark the party-line option and walk away without knowing that they have failed to vote for president. In North and South Carolina, even though they are party line states, DREs perform quite well relative to other systems in those states. If, however, a county in another party line state were to have this option selected and it was not so diligent in its notification to voters of the fact, a few voters may not notice that they are failing to vote for president. This could account for the poor performance of DREs in some party line states.

We attempted to collect data on which counties used which options, but this proved exceedingly difficult. Our first approach was to contact the SoS of each party line state which had DRE systems and ask which options they used. For North Carolina and South Carolina, this is a simple matter since the option selection is a matter of state law. Other states do not keep detailed records of the options used by specific county voting machines. States tend to set guidelines for what systems are acceptable, and leave the options to the counties, who are the primary administrators of elections. Furthermore, the options for configuration of DRE systems, such as Microvote's party line options, are not well-known, even to the administrators. While most SoS offices, such as Indiana, told us that they did not keep such records and we would have to contact the counties, many county officials did not even know that such options existed. After explaining in depth the configuration options, administrators tended to tentatively claim to use the one-button-only option, although we believe that many counties may not know how their systems are configured. As a result,

we do not have much confidence in any such data that might be collected. Consequently, we cannot test whether or not this is responsible for the poor performance of DREs in some party-line states, but it is a possible explanation.

However, given that configuration options must be selected at the time of purchase, and can only be changed by repair personnel, it seems unlikely that an administrator might simply accidentally flip a switch and fail to inform party line voters of this change. Also, were this to be the case, there would probably be at least a precinct or two for which the residual vote rate was at least 30% (if a DRE was configured for the president-excluded party line vote option and nobody knew it, most likely everyone who chooses a party line vote will fail to cast a presidential vote). We believe that the greater issue is one of voter notification and instruction. If a county chooses to use the president-excluded option of party line voting, and if the instructions are buried in the text of the voter's instructions, there will be at least a few people who will miss it, and will fail to cast a presidential vote. The more clear a voter's instructions are, the more successful we expect the implementation of a new system to be.

There is still another possible explanation for the high residual vote rates for DREs in some party line states. New Mexico is a party line state and the DREs used in some of its counties have relatively high residual vote rates. We could find no administrative or technical reason for these anomalies, but several sources, including the Director of Elections in the New Mexico Secretary of States' office, suggested that some counties in New Mexico had a large number of citizens, particularly Native Americans, who were more interested in local elections than in national elections, and showed up to the polls to cast their votes in local elections without any intention of voting for president. Although this may be true, we could not find any clear-cut relationship between residual vote and the percentage of Native Americans in a county.

These comments suggest that we still need to learn a great deal more about how voting systems are implemented and how they are used. In the first part of this section we cited some evidence suggesting important differences across subtypes of voting systems (e.g., Datavote versus Votomatic; precinct count versus central count optical scan), but we noted that the national data are not detailed enough to compare these systems. In the second part of this section we discussed some explanations for the higher rates of residual votes for DREs in party line states. The first explanation focused on election administration. The second upon the idiosyncratic voting habits of a particular voting population. Despite the possible importance of these factors, we have very little detailed data on election administration and on the habits of specific voting populations. There are clearly many areas where more data collection and research is needed.

X. TWO CALIFORNIA CASE STUDIES

In this section, we use data from the California Statewide Database[58] and other sources to study two examples of changes in technology. We do not claim that these examples are representative, but we believe that they illustrate what can be done with new technologies.

Fresno County used Votomatic style punchcards in 1996 and replaced them in 2000 with a precinct count optical scan system. Riverside County used a central count optical scan system in 1996 for all its voters and changed to a touch screen DRE for election day and early voting in 2000. It retained the central count optical scan system for absentee voters. These two cases provide us with a chance to see how optical scan and DRE systems perform in two large counties with large cities and some rural hinterlands.

First consider Fresno County. Figure 6 uses information from the California Statewide Database to plot 2000 presidential residual vote versus 1996 presidential residual vote for 124 Census tracts. Note that the vertical axis for the 2000 vote goes from zero to three percent while the horizontal axis for the 1996 vote goes from zero to eight percent. There is a significant reduction in residual votes in every single tract. In fact, the average residual vote went from 3.35% in 1996 to 0.70% in 2000—a movement from an "unacceptable" voting system in the National Commission's reckoning to a "good" one.

Now consider Riverside County which used a central count optical scan system in 1996. As we have seen, optical scan systems typically work relatively well. In fact, in 1996 the average presidential residual vote rate was 1.21%—an "adequate" performance according to the Carter/Ford Commission. In 2000, early voters and election day voters used a touch screen system, and the residual vote rate was .59%—a "good" performance. The 37% of the Riverside voters who voted absentee still used the central count optical scan system. This fact provides us with the opportunity to compare voters in the same precinct who voted using DREs with those who used optical scan. Figure 7 shows the improvements that occurred. Because precincts are much smaller than tracts, this figure is noisier than Figure 6, but the same result is apparent.[59] The horizontal axis for the optical scan system is elongated compared to the vertical one for the electronic system because of the better performance of the electronic system. Whereas the residual vote rate for the optical scan system was 1.21% in 2000 (the same as it was in 1996), the residual vote rate was only half that for the electronic system.

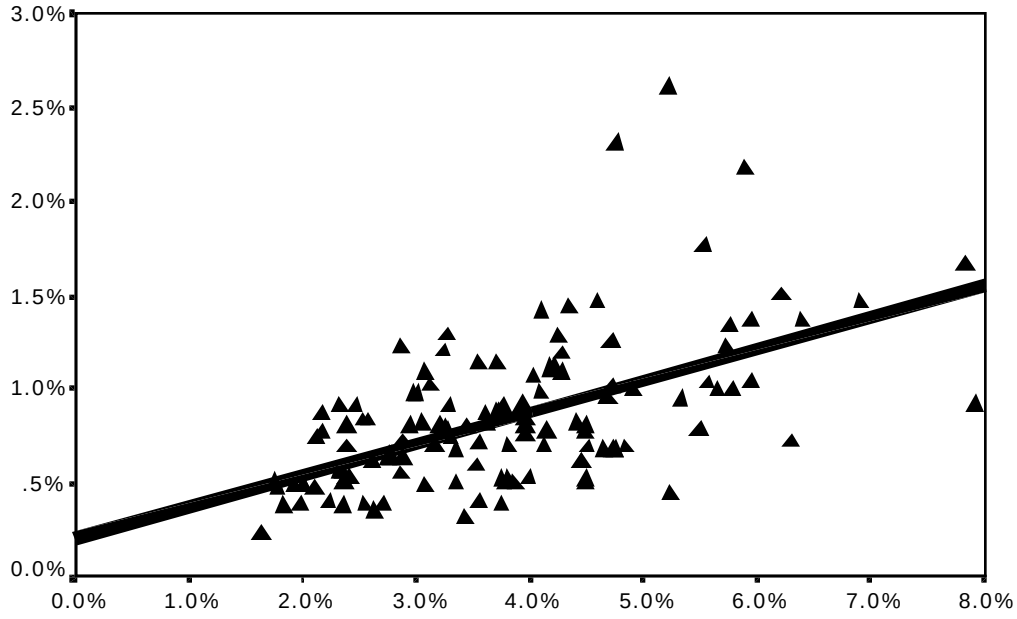
[58] The California Statewide Database is the redistricting database for the state of California which is made available through the Institute of Governmental Studies of the University of California, Berkeley. In 2001, these data are being used for both the state legislative and local redistricting, mandated by law after the next census.

[59] In order to smooth out this figure, we average the absentee ballots and residual votes for each precinct with two adjacent precincts so that the number of absentee ballots was roughly the same as the number of election day voters.

Figure 6

Presidential Residual Vote Percentage Change in

Census Tracts in Fresno County--1996-2000

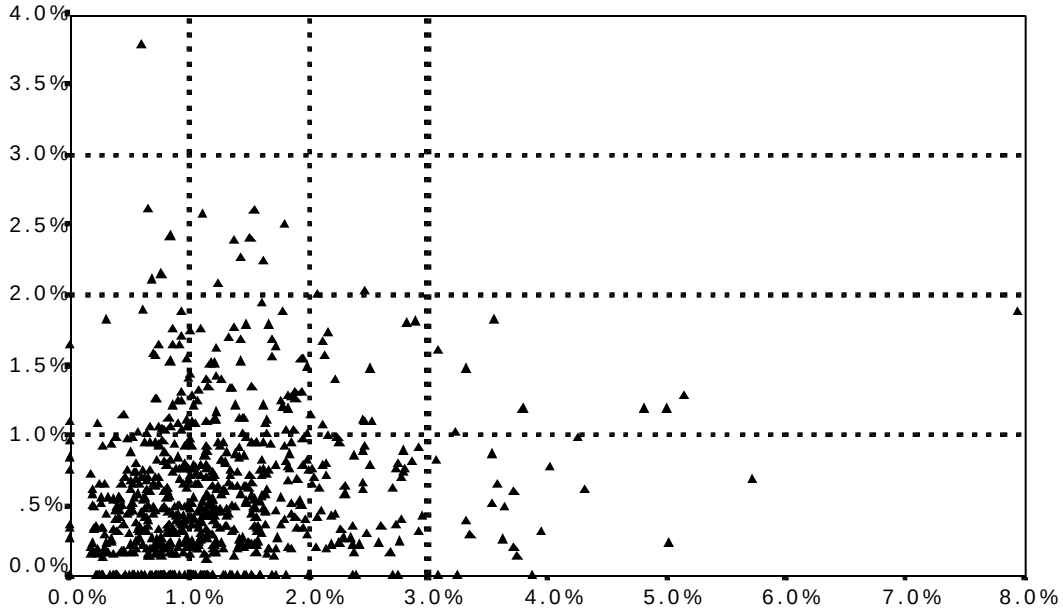


1996 Presidential Residual Vote Using Punchcards

Figure 7

Presidential Residual Vote for DRE

Versus Optical Scan Voting in Riverside



Residual Vote--Absentee Voting with Optical Scan

XI. COUNTY ACQUISITION DECISIONS AND SELECTION BIAS

Perhaps the greatest problem with evaluating voting systems is that the technologies continue to change. These changes present special problems for observational studies such as this one, and they present some subtle, yet important, issues that arise from the fact that voting systems are not randomly assigned to counties—they are steadily being procured by counties through a process that includes at least some consideration of what system will perform best in the county. This process leads to what statisticians call "selection bias" which can confound statistical inferences.

Technological change is behind this process. Paper ballots have been and will continue to be available for a long time, but mechanical lever machines are no longer being produced. The design of most punchcard systems dates from the 1960's and 1970's. Optical scan systems are newer, and precinct scanning systems date from the mid-1980's. DREs, especially touch screen DREs, are quite new. As noted above, in the past forty years there has movement away from paper ballots and lever systems to punchcard systems, and now from all three to optical scan and DREs. Two of the major determinants of these changes have been costs and the speed with which ballots could be counted. Paper ballots are relatively inexpensive, but they are hard to count reliably. Mechanical lever machines provide instant tabulations, but they are expensive, bulky, and relatively hard to maintain. Punchcards systems were inexpensive and portable compared to mechanical lever machines, and they provided a quick way to count and recount ballots.

Because the performance of voting systems with respect to overvotes and undervotes has only been a side consideration for most counties, producing an impetus for change only when the performance was clearly unacceptable, it is not surprising that there still remains a great range in the residual vote performance of voting systems. But it also seems likely that voting systems have retreated to those counties where their residual vote performance was acceptable when stacked up against their other characteristics. For mechanical lever machines that are quite expensive, hard to maintain, and hard to store, therefore, it is not surprising that their residual vote performance had to be quite good. After all, they had to have some comparative advantage. For punchcard systems, it has meant that low cost could excuse poor residual vote performance, although punchcards have probably been banished from those places where they performed worst. For paper ballots it has meant that they have found a niche in small rural counties where the rapid vote counting capabilities of the two other systems were not so important. In fact, paper ballots are now used in counties (outside of the township states) whose median number of voters is about one-fifth to one-tenth that for all other systems, and for which the number of voters per precinct is about forty percent of that for other systems. In these small counties with about 200 voters per precinct (twenty to thirty people per hour during a voting day), purchasing expensive equipment makes no sense and votes can be counted relatively quickly and carefully without it. Moreover, it is not surprising that many of the townships in township states still use paper ballots—after all, townships are small localities that can best utilize paper ballots.

Finally, both precinct count optical scan systems and DREs are relatively new systems, and it will take time for voters and election officials to become familiar with them. We believe that both have already demonstrated their ability to work well as shown by the case studies of Fresno and Riverside Counties. But they still need further development, and we believe that human factors engineering and experiments in field settings would go a long way towards improving them and providing a better evaluation of them. We hope that the federal government, state governments, foundations, and the vendors who make these systems take up this challenge and create systems that can all be called "good" by the standards of the National Commission on Federal Election Reform.

SUMMARY OF FINDINGS AND RECOMMENDATIONS

Goal and Purpose of the Study

The goal of this study was to determine which voting systems do the best job of recording and tabulating votes. Our primary measure of performance was the residual vote percentage—the percentage of ballots in a jurisdiction for which no presidential votes were recorded.

We considered residual votes in U.S. counties in the 2000 general election. Counties are the units that, in most cases, manage elections most directly and determine voting systems. Most of our emphasis is on the presidential election because we are skeptical about what residual votes mean for races down the ticket. We have used data for 2000 because it the only information of suitable quality and, luckily, it also provides the most up-to-date picture of the relative performance of voting systems.

Data Quality

Our dataset includes information on the type of voting system, residual votes computed from turnout and votes for candidates, and demographic and political characteristics for 2219 counties. These counties constitute almost three-quarters of the counties in the United States, three-quarters of the states, three-quarters of the population, and three-quarters of the presidential votes cast in the 2000 presidential election. Thus, these data provide a highly representative sample of the voting systems and voting data across the country.

We had to make an extensive investment in collecting and cleaning these data before they were acceptable to us. **Consequently, one of our major concerns is the quality of these data.** Specifically, we have the following concerns:

County voting systems. We are relatively confident in our classifications of county voting systems, and for every county included in the data, we have at least two sources agreeing on the system used. Nevertheless, we still believe that there are unavoidable errors given the difficulty of data collection and the erratic record-keeping of some counties.

Turnout data. We have substantial skepticism about the quality of the turnout data that we have collected. Moreover, we worry that there may be a complicated interaction between state laws for reporting turnout and voting systems that may affect the total ballot/turnout figures that we use to compute residual vote.

Demographic and Political Data. Although we have confidence in the quality of the Census demographic data, we worry that the description of counties is still incomplete because of the limited amount of information from the Census. Specifically, it would be very useful to have much more information about election administration and political factors in each county.

Our concerns with the data led us to reject the idea of using data for earlier years even though we had obtained such data from Election Data Services. After finding many errors and oddities in the 2000 data, we became skeptical of the quality of the 1996 data. We have no reason to believe that these data would be any more reliable than the error-ridden 2000 information. And auditing the 1996 data with the same care with which we audited the 2000 data would be impossible because we found that numerous county officials were unsure about which voting technology they used in 1996. We concluded that it would be a serious misjudgment to use data that are likely to be so error-prone.

Using Residual Votes as the Measure of Performance

Our report focuses narrowly on one dimension of performance: the degree to which voting systems record votes for every voter who comes to the polls. This measure, however, combines *intentional* undervotes which should not be attributed to poor performance by voting systems with *unintentional* undervotes and overvotes which could be the result of poor performance by these systems. We present evidence to show that a majority of presidential residual votes are unintentional undervotes and overvotes so that a residual vote measure for presidential elections, with proper statistical controls, can provide information about the performance of voting systems.

We are much less sure about the adequacy of this measure for senatorial and gubernatorial elections because so many of the residual votes may be intentional undervotes whose numbers may differ substantially from one jurisdiction to another depending upon the closeness of the election. The major complication is that although *some* senatorial and gubernatorial residual votes are undoubtedly unintentional undervotes or overvotes caused by voters' experiencing problems with voting systems, a significant proportion is simply the result of voters intentionally deciding not to vote in the race. As a result, a fair comparison of average residual votes across systems must confront the difficult task of finding a reliable way to determine how much of the senatorial residual vote is unintentional and how much is intentional.

Methodological Issues

All in all, our claims are quite modest because of the difficulties of drawing conclusions from non-experimental, observational, data. **We make modest claims because there are four major problems in assessing voting systems with national data:**

- **Poor data and measures**
- **Difficulties statistically controlling for other factors that affect performance**
- **Older systems remain in situations where they work so comparisons may be biased**
- **Gross categorizations of diverse and sometimes rapidly changing technologies**

These problems pose significant challenges to researchers who want to draw conclusions from national data.

Presidential Residual Votes and Voting Systems

We find that direct record electronic, lever machines, optical scan, and paper ballots <u>all</u> produce significantly fewer residual votes, between 1/2% to 1% less on average, than punchcards. This result remains true when the data are analyzed in different ways and with statistical controls.

Using the four point scale devised by the National Commission on Federal Election Reform which uses the gradations of "good," "adequate," "worrying," and "unacceptable" performance, we find that all systems have a substantial number of unacceptable and worrying implementations, but punchcard systems are typically in the bottom two categories while all other voting systems are typically in the top two categories.[60] The available data are insufficient for saying which one of the non-punchcard systems performs best.

We find, for example, that in the 100 largest counties in the United States comprising about 40 percent of the U.S. population, optical scan and electronic systems can be given the highest rating of "good" on the four point scale devised by the National Commission on Federal Election Reform (the Carter/Ford Commission). Lever machines are "adequate" by this definition and punchcards are "worrying". Optical scan systems do better than any other system when the simple average of presidential residual votes is considered for all U.S. counties, but this advantage disappears when we introduce basic

[60] These statements provide our summary judgments based upon all the analysis in this report. The statements average over all of the analysis for each voting system. In some specific analyses, especially those involving averages over all counties, we have found that one or two of the non-punchcard systems tend towards the second and third categories instead of the first and second, but this is balanced by substantially better performance in other analyses.

statistical controls. In fact, lever machines seem to do best when controls are introduced. Optical scan, DREs, and paper ballots are tied for second place. Paper ballots do well in small counties.

We also find that DREs and paper ballots are substantially less prone to what the Cal-Tech/MIT team call "very high residual vote rates" (5% residual vote or higher) than any other system. Only 4.0% of counties using paper ballots and 4.5% of counties with DREs perform this poorly—the corresponding figures for other systems are 9.9% of counties with lever machines, 9.0% of those with punchcards, and 7.5% of those with optical scan systems.

In summary, paper ballots typically do well, but they are only used in small counties (the largest has a population less than 60000 and the average population is 6100 people) and the data tell us virtually nothing about what would happen if they were used in larger counties. Lever systems are competitive with optical scan systems and DREs across all counties but they seem prone to very poor performance and they do not do well in large counties. Optical scan and DREs appear to dominate all other systems in terms of overall performance across all counties and especially in large counties. DREs do not do as well as optical scan systems in smaller counties, but optical scan systems seem slightly more prone to very poor performance than DREs. The data suggest that we need to know much more about all of these systems because about a third of the population and more than half the counties in America have systems that are "worrying" or "unacceptable."

The available evidence does not provide any substantiation for Cal-Tech/MIT's claim that DREs might be harder to use than other systems. In fact, we present evidence suggesting that DREs pose fewer problems for poorly educated voters than all other systems except paper ballots. Moreover, punchcards appear to be very error prone for everyone. Even voters with higher levels of education seem to have trouble with them.

Based upon our own analysis of data for the 2000 election and our concerns with these problems, we advise against the rush to judgment of the Cal-Tech/MIT report with respect to the relative performance of non-punchcard systems, and we agree with the Commission on Federal Electoral Reform and the National Task Force on Election Reform that much more testing of the available systems is needed before making any final judgments about the suitability of one system over another.

Senatorial and Gubernatorial Residual Votes and Voting Systems

We are skeptical about using senatorial or gubernatorial residual votes to evaluate voting systems, but we have undertaken an analysis to demonstrate the complications. Our analysis of senatorial residual votes finds that there is some evidence that lever and punchcard systems perform poorly in recording senatorial votes and that other systems perform better, but there is no consistent finding about the relative performance of optical scan, DRE, and paper ballots. Most importantly, our results also strongly suggest that we should not use senatorial or gubernatorial residual votes as a standard for evaluating these systems until we know a lot more about intentional undervoting than we currently know.

Caveats

One of our major findings is that we need much more data and information before we can provide a fully detailed assessment of voting systems. We do not know enough about the details of voting systems and their implementations to make careful distinctions among systems with respect to their residual vote performance, and we do not have enough information about how voting systems perform with respect to criteria other than residual vote performance. Data for California counties, for example, suggest that Datavote punchcards perform better than Votomatic-style systems, and the national data suggest that precinct count optical scan performs better than central count optical scan. But the national data on voting systems is not precise enough to make it possible to test these possibilities carefully, partly because there is limited information about the subtypes of voting systems. In addition, our

investigations suggest that the specific implementations of voting systems, the quality of election administration, and the idiosyncratic characteristics of voting populations matter a great deal for residual vote performance, but the available data do not allow us to study these factors.

Furthermore, residual vote percentage is just one measure of the performance of voting systems. Features of voting systems other than their ability to record and tabulate votes should also be considered. Touch screen DREs, some of the earlier pushbutton DREs, and optical systems that scan ballots in the precinct have the advantage that voters can be given feedback that "checks their work" before they submit their ballot. Touch screen DREs have the potential for making it possible for language minorities to choose their ballot language in the voting booth and for those with disabilities to vote by listening to a recording. In addition, voting systems differ in their costs, their portability and their ease of set-up. All these factors should be taken into account when choosing a system.

Recommendations

We recommend the following with regard to voting systems in the United States:

1. Concerted efforts should be made to move away from Votomatic-style punchcard systems to other, more appropriate ones. In most cases this will mean either optical scan or electronic systems.

2. States should implement uniform reporting standards for county reporting of undervotes and overvotes. The Office of Election Administration of the Federal Election Commission should take the lead in developing these standards. At the very least, information about undervotes and overvotes should be reported by precincts and by absentee, early voting, and election day voting after every election. States should also keep up-to-date records on voting systems in each county or township.

3. More effort should be put into human factors engineering to find out what features of voting systems cause unintentional undervotes and overvotes.

4. Voting systems should be studied using experimental methods which vary their features to see which ones have the most impact on their performance.

In addition, we recommend the following regarding research on voting systems:

1. More and better data should be collected at the county and precinct level on:

   - The types of voting systems and their implementation,
   - The process of voter education, and
   - The level of staffing and resources devoted to election administration.

2. A national program of experimentation with voting systems should be undertaken that involve experts in engineering, human factors, psychology, political science, and economics who would use the best available experimental methods to study voting systems.

Appendix 1: Additional Peculiarities in the Data Set

The remaining peculiarities of the data set involve the residual vote percentages. Nine states have one or two counties with negative residual vote percentages (Alabama, Colorado, Louisiana, Nebraska, New Jersey, North Carolina, Rhode Island, South Carolina, and Washington). Three states, Indiana (7), Kansas (11), and Kentucky (12), have more than one county with negative residual vote percentages. As mentioned in the main body, all of these 42 counties were eliminated from the dataset although data was run with and without them to determine their impacts. Three negative residual rates in Kansas counties using paper ballots that were greater than -.15 did have a substantial effect on our estimates for paper ballots, but we believe that these values are so fantastic that they can be safely eliminated.

There is, of course, the question of why some residual rates are less than zero when, by definition, they should be greater than or equal to zero. They must be the result of errors. We believe they come from the following kinds of mistakes: absentee votes and election day votes may be summed to get votes cast, but ignored in the turnout figures because they never actually appeared at the polling places, so the figures reported for total number of voters showing up may be misleadingly low. Or the polling place roster of signatures is scanned into a computer to generate a list of voters who turned out, but if there is, for some reason, an error in reading a name, that name will be ignored, and again, the figures for turnout may be misleadingly low. Or when states purge voter rolls, there may be confusion at the polls, and voters may be given provisional ballots without their names being on the list. Since a negative residual vote makes no sense, we experimented with excluding them all from our analysis, but it made very little difference in our substantive analysis because the same problem affects a number of counties reporting positive residual votes as well. Essentially, an undetermined number of counties reported a lower residual vote than they really had. Some of these counties report positive residual votes, and some report negative residual votes if their true residual votes were few enough. Thus, since the problem exists throughout the data set, it may not make much sense to arbitrarily exclude negative residual votes from our analysis, although again, it made little difference in our substantive analysis when we did so because the important figures are the differences between residual votes with different voting systems, and if this bias effects all counties equally, those differences are unaffected.

In addition, seventeen counties, seven of them in Alabama, and the rest in Colorado, Georgia (2), Indiana, Iowa (2), Kentucky, Nebraska, North Carolina, and Ohio, reported an exactly zero residual vote percentages. Such reports may be the result of the problems described above, but they are more likely the result of the counties mis-reporting the number of votes cast for the presidential candidates as their turnout/ballots which would lead to zero percent residual vote by mathematical necessity.

References

Achen, Christoper H., and W. Phillips Shively. 1995. Cross-Level Inference. Chicago: University of Chicago Press.

Analytic Systems, Inc. 1974a. Describe, Analyze and Compare the Currently Available Methods of Vote Counting Equipment and to Make Appropriate Recommendations: Final Report. Washington, DC: U.S. General Accounting Office.

Analytic Systems, Inc. 1974b. Experimental Voting System Supplement. Washington, DC: U.S. General Accounting Office.

Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. The American Political Science Review 88: 829-838.

Bain, Henry M., and Donald S. Hecock. 1957. Ballot Position and Voter Choice. Detroit: Wayne State University Press.

Barnes, Samuel, et al. 1962. "The German Party System and the 1961 Federal Election." American Political Science Review 56, 4 (December): 899-914.

Burnham, David. 1985a. New York Times. "Computerized Systems for Voting Seen as Vulnerable to Tampering," July 29.

Burnham, David. 1985b. New York Times. "Votes by Computer: Some See Problems," August 21.

Burnham, David. 1986. New York Times. "Texas Looks Into Reports of Vote Fraud." September 23.

Burnham, Walter Dean. 1965. "The Changing Shape of the American Political Universe." American Political Science Review 59 (March): 7-28.

Cal-Tech/MIT. 2001. Voting: What Is, What Could Be?

Cal-Tech/MIT Voting Technology Project, "A Preliminary Assessment of the Reliability of Existing Voting Equipment," Version 1: February 1, 2001.

Cal-Tech/MIT Voting Technology Project, "Residual Votes Attributable to Technology: An Assessment of the Reliability of Existing Voting Equipment," Version 2: March 30, 2001. Available at

Carey, Jane and Andrew Carey. 1957. "The Varying Seasons of Italian Politics, 1956-57." Political Science Quarterly 72, 2 (June): 200-223.

Carey, Jane and Andrew Carey. 1958. "The Italian Elections of 1958—Unstable Stability in an Unstable World." Political Science Quarterly 73, 4 (December): 566-589.

Cox, Gary W. & Michael Munger. 1989. "Closeness, Expenditures, and Turnout in the 1982 U.S. House Elections." American Political Science Review 83 (March): 217-231.

Crain, Mark, Donald R. Leavens, and Lynn Abbot. 1987. "Voting and Not Voting at the Same Time." Public Choice 53: 221-9.

Darcy, R. 1986. "Position Effects with Party Column Ballots." Western Political Quarterly 39 (December): 648-62.

Darcy, R. and Anne Schneider. 1989. "Confusing Ballots, Roll-off, and the Black Vote." Western Political Quarterly 347-364.

Dubois, Philip L. 1979. "Voter Turnout in State Judicial Elections: An Analysis of the Tail on the Electoral Kite." Journal of Politics. 41: 865-887.

Dugger, Ronnie. 1988. "Annals of Democracy." The New Yorker. November 7. 40-108.

Election Data Services. 1981. Voting Systems, Vendors 81: Equipment Available on the Market. Washington, DC: National Clearinghouse on Election Administration.

Florida Governor's Select Task Force on Election Procedures, Standards, and Technology. March, 2001. Revitalizing Democracy in Florida, Collins Center for Public Policy, Tallahassee, Florida.

Garber, Marie. 1998. Innovations in Election Administration, Volume 17: Acquiring Election Systems and Equipment. Washington, DC: Office of Election Administration.

Goodman, Leo. 1953. "Ecological Regression and the Behavior of Individuals." American Sociological Review 18: 663-64.

Goodman, Leo. 1959. "Some Alternatives to Ecological Correlation." American Journal of Sociology 64: 610-25.

Harris, Joseph P. 1934. Election Administration in the United States. The Brookings Institution, Washington, D.C.

Kelley, Jonathan, and Ian McAllister. 1984. "Ballot Paper Cues and the Vote in Australia and Britain: Alphabetic Voting, Sex, and Title." Public Opinion Quarterly, 48 (Summer): 452-66.

Kim, Jae-on and B. C. Koh. 1972. "Electoral Behavior and Social Development in South Korea: An Aggregate Data Analysis of Presidential Elections." The Journal of Politics, 34, 3 (August): 825-859.

King, Gary. 1997. A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton: Princeton University Press.

Knack, Stephen and Martha Kropf. 2001. "Roll Off at the Top of the Ballot: Intentional Undervoting in American Presidential Elections." Working Paper.

Moloney, Martha. 1975. Mechanized Vote Recording: A Survey. Frankfort: Legislative Research Commission.

Montgomery, Michael. 1985. "Voting Systems and Disenfranchisement." Election Politics, 2 (Summer): 16-19.

National Clearinghouse of Election Administration. 1984. Voting System Standards. Washington, DC: National Clearinghouse on Election Administration.

National Commission on Federal Election Reform. August 2001. To Assure Pride and Confidence in the Electoral Process. Miller Center of Public Affairs, University of Virginia.

National Scientific Corporation. 1977. Voting Systems, Volume 1: Recommended Procurement Procedures and a Review of Current Equipment. Washington, DC: Federal Election Commission Clearinghouse on Election Administration.

National Task Force Report on Election Reform. July, 2001. Election 2000: Review and Recommendations by the Nation's Elections Administrators. The Election Center, 12543 Westella, Suite 100, Houston, Texas.

Roth, Susan King. 1998. "Disenfranchised by Design: Voting Systems and the Election Process." Information Design Journal. Volume 9, Number 1.

Rusk, Jerrold G. 1970. "The Effect of the Australian Ballot Reform on Split Ticket Voting: 1876-1908," American Political Science Review. 64 (December): 1220-1238.

Saltman, Roy. 1975. Effective Use of Computing Technology in Vote-Tallying. Washington, DC: U.S. Department of Commerce.

Saltman, Roy. 1988. Accuracy, integrity, and security in computerized vote-tallying. Gaithersburg, MD: U.S. Department of Commerce National Bureau of Standards.

Stiefbold, Rodney. 1965. "The Significance of Void Ballots in West German Elections." American Political Science Review, 59, 2 (June): 391-407.

Strong, Donald. 1948. "The Rise of Negro Voting in Texas." American Political Science Review, 42, 3 (June): 510-522.

Thomas, Norman. 1968. "Voting Machines and Voter Participation in Four Michigan Constitutional Revision Referenda." Western Political Quarterly, 21 (September): 409-19.

Vanderleeuw, James, and Richard Engstrom. 1987. "Race, Referendums, and Roll-Off: A Research Note." Journal of Politics, 49, 4 (November): 1081-1092.

Walker, Jack. 1966. “Ballot Forms and Voter Fatigue: An Analysis of the Office Block and Party Columns Ballots.” Midwest Journal of Political Science, 10 (November): 448-64.

Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane, and Michael C., Herron. Forthcoming. “The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida,” American Political Science Review.