Reliability of the Uncertified Ballots in the 2000 Presidential Election in Florida

Kirk WOLTER, Diana JERGOVIC, Whitney MOORE, Joe MURPHY, and Colm O'MUIRCHEARTAIGH

Following one of the most closely contested elections in American history, a group of the nation's largest media organizations retained the National Opinion Research Center (NORC) to conduct an in-depth inventory of all uncounted ballots from the 2000 presidential race in Florida. This article describes the planning and implementation of the project and its resulting databases. The State of Florida employed three major voting systems for the election: two systems based on punch cards, and various optical scanning systems. We analyze and present statistics regarding the reliability of the various voting systems. Although none are perfect, we generally found optical scanning to be superior to the punch card systems. We summarize analyses of project databases conducted by the Media Group with respect to new vote totals that hypothetically might have been achieved had the U.S. Supreme Court not stopped the vote counting. We offer recommendations to policy makers for future elections.

KEY WORDS: Agreement; Bush; Coding; Consistency; Gore; Multilevel models.

1. INTRODUCTION

The presidential election of 2000 was among the closest and most interesting elections in American history. In the state of Florida, 6,138,120 ballots were cast. At the time the U.S. Supreme Court stopped the vote counting, 175,010 ballots (or approximately 2.9%) remained uncertified for the presidential race, including 61,190 undervotes and 113,820 overvotes. The uncertified rate is comparable to rates in the presidential elections of 1996 and 1992, with 2.5% and 2.3% of the ballots uncertified, respectively. However, the 2000 election was so close that every ballot could potentially have made the difference.

Following the 36-day, presidential-election crisis, a group of the nation's most-respected media organizations (*The New York Times*; *The Washington Post*; *The Wall Street Journal*; CNN; The Associated Press; The Tribune Publishing Company, represented by *The Los Angeles Times*, *The Chicago Tribune*, and *The Orlando Sentinel*; Cox News Service, representing *The Palm Beach Post*; and *The St. Petersburg Times*) hired the National Opinion Research Center (NORC) at the University of Chicago to conduct a thorough review of Florida's uncertified ballots. This team recognized the historical significance of the uncertified ballots, and worked together to create a definitive historical archive of these ballots. This formidable task was enabled by Florida's sunshine law (P.L. 101.572), which authorizes anyone to view ballots from a state election.

The Media Group and NORC adopted two main analytic goals. The Media Group took as its principal goal the determination of which uncertified ballots could be assigned as votes to specific presidential candidates, and consequently of which candidate might have won the state of Florida given alternative voting counting scenarios or standards. NORC's analytic goal was to measure and compare the reliability of the various voting systems in use in Florida, with the aim of providing elections officials with a base of information to guide the improvement of future elections. On November 12, 2001, NORC released the archive to the public via its Web site (NORC.ORG) and the Media Group published or aired the results of their analysis of vote totals. This article mainly contains the results of NORC's analysis of reliability.

The state of Florida is divided into 67 counties, which used five different voting systems. The majority of counties (41) used a system by which voters used a specified pen or pencil to fill in arrows or ovals to select candidates. An optical scanner would read the ballots and register votes for the candidates. Fifteen counties used a Votomatic punch card voting system. This system required the voter to insert a computer punch card, containing many prescored chad (a small area of approximately 1/16 inch in diameter with a perforated border), into a device and then to use a stylus to punch out the chad for the selected candidate. A machine—which recorded the number of chads punched out for each candidate—counted the votes. Nine counties used a Datavote system, which was similar to Votomatic technology in that voters were required to insert a computer punch card into a device. For the Datavote system, however, the voter aligned a mechanical punch tool with the candidate of choice and punched

Kirk Wolter is currently Director, Interdisciplinary Research Institute for Survey Science, and Professor of Statistics, Iowa State University (E-mail: wolter@iastate.edu). Diana Jergovic is Director, Operations and Planning, Office of the Vice-President for Research and for Argonne National Laboratory, University of Chicago. Whitney Moore is Sampling Statistician, National Opinion Research Center. Joe Murphy is Research Survey Methodologist, RTI International. Colm O'Muircheartaigh is Vice President, Statistics and Methodology, National Opinion Research Center, and Professor, Irving B. Harris Graduate School of Public Policy Studies, University of Chicago. All authors were with NORC at the time of the Florida Ballots Project. The authors thank Jim Davis for collaboration in the development of Section 3 and for general discussion and analysis of coder reliability; Barbara Bailar for collaboration in the development of Section 4 and for general discussion and analysis of coder reliability; and Anirban Basu, Anthony Bryk, and Pamela Campanelli for helpful discussions on the application of hierarchical models to this problem. We also thank the referees for helpful comments.

 DOI: 10.1198/0003130031144

Table 1. Florida Voting Systems by Number of Counties, Number of Uncertified Ballots Examined by NORC, and Total Number of Ballots

Voting System	Number of Counties	Undervotes	Undervotes as a Percent of Total Certified and Uncertified Ballots	Overvotes	Overvotes as a Percent of Total Certified and Uncertified Ballots	Total Uncertified Ballots	Uncertified Ballots as a Percent of Total Certified and Uncertified Ballots	Total Certified and Uncertified Ballots
Votomatic (punch card with chad)	15	53,215	1.5	84,822	2.3	138,037	3.8	3,642,160
Datavote (punch card without chad)	9	771	0.6	4,427	3.2	5,198	3.7	138,869
Optical scan	41	7,204	0.3	24,571	1.0	31,775	1.3	2,357,091
Lever	1	†						
Paper	1	‡						
Total	67	61,190	1.0	113,820	1.9	175,010	2.9	6,138,120

† Ballot totals for the lever county are included in totals for the Datavote counties.
‡ Ballot totals for the paper county are included in the totals for the optical scan counties.

a hole into the ballot. There were no prescored chads. A machine then read the punch cards and counted the votes. One county used a lever voting system—a system by which voters did not use ballots. A large voting apparatus simply tallied votes. Finally, one county used paper ballots on which voters used any pen or pencil to indicate selected candidates by marking an X in a box next to the candidate's name. Paper ballots were counted by hand.

None of the voting systems are perfect and uncertified ballots can result from mistakes by the voters, errors by the counting system (machine or canvassing board), or intentional actions of the voters. An undervoted ballot means the machine count (or, in some cases, the subsequent hand count) records no vote for president, such as when a voter intends not to vote in the presidential race. Undervotes from Votomatic technology may also occur when a chad is not completely punched out or when a pen is used to mark a chad (instead of a stylus used to punch out a chad). Undervotes from optical scan systems result when something other than the specified pen or pencil is used to vote. Because scanners can read only certain markings, using the designated writing implement is crucial to casting a successful vote. Optical scan undervotes can also result when marks are made outside of the designated oval or arrow or when the marks made within the oval or arrow are not dark or complete enough. Datavote and lever systems can produce only intentional undervotes. However, absentee ballots in these counties can be undervoted if a pen is used to indicate selected candidates. Finally, paper voting systems produce only intentional undervotes. The critical issue with an undervoted ballot is that the machine (or in some cases the county canvassing board) did not detect a vote for the presidential race.

Ballots are overvoted when more than one candidate is selected for the presidential race. Intentional and unintentional overvoting is possible with all voting systems except the lever system, which deactivates the first candidate if a second candidate is selected in a race. Votomatic and Datavote ballots are subject to overvoting if a correction is attempted by simply punching for another candidate after the first candidate has been indicated. Optical scan and paper ballots are subject to overvoting when a correction is attempted by crossing out or otherwise negating one selection and indicating another selection on the same ballot. The optical scanning machines cannot determine which mark is the intended vote. Paper ballots are all reviewed by humans, and thus present fewer ballot errors of this type. Finally, the misuse of the write-in section of the ballot may create an overvote. A voter who voted for a candidate and then wrote in that candidate's name or the name of another candidate created an overvote that could not be certified for any candidate. The critical issue with an overvoted ballot is that the machine (or county canvassing board) detected more than one vote in the presidential race.

Rates of ballot error differ by type of voting system. While only 22% of Florida's counties used Votomatic technology, approximately 79% of the uncertified ballots were from Votomatic counties and 3.8% of Votomatic ballots were uncertified. Optical-scan counties (61% of all Florida counties) and one paper county produced approximately 18% of the uncertified ballots, and 1.3% of their total ballots were uncertified. Datavote counties (13% of all Florida counties) and one lever county accounted for approximately 1% of the uncertified ballots, and 3.7% of their total ballots were uncertified. Votomatic had an undervote error rate of 1.5% versus 0.6% and 0.3% for Datavote and optical scan, respectively. Details appear in Table 1. These simple statistics suggest that optical-scan systems may present fewer problems to voters than punch card systems. (Optical scanning and other high-tech solutions, however, are still no panacea, as was demonstrated recently in the 2002 Democratic primary election in Florida.)

In what follows, we present specific evidence about the reliability of hand-counting the various types of ballots, given that the machine classified the ballots as undervotes or overvotes.

2. METHODOLOGY

To pursue our analytical goals, we devised teams of three workers (called coders) who reviewed and recorded the nature of the markings on every one of the uncertified ballots. This type of information allowed the Media Group to examine vote totals given a variety of suggested vote counting standards, and NORC to study the variability in the coding or the reliability of the ballot systems. Three coders per ballot also provided a

Table 2. Distribution of Florida Coders and of US Adults in the 2000 General Social Survey (GSS)

Coder characteristics	Coders (in percent)	GSS 2000 (in percent)
Sex	$N = 152$*	$N = 2{,}817$
Male	29.6	44
Female	70.4	56
Total	100.0	100
Age (recoded from date of birth)	$N = 148$	$N = 2{,}832$
63+	23.6	16
53–62	25.0	12
41–52	26.4	24
20–40	25.0	45
18–19	0.0	2
Total	100.0	99
Education	$N = 152$	$N = 2{,}799$
Graduate degree	9.9	8
Bachelor's degree	23.7	16
High school graduate	65.1	61
Less than high school	1.3	16
Total	100.0	101
Family Income	$N = 150$	$N = 2{,}456$
$100,000+	6.0	7
$75,000–100,000	9.3	9
$50,000–75,000	17.3	18
$25,000–50,000	34.7	29
Less than $25,000	32.7	36
Total	100.0	99
Race/Ethnicity	$N = 150$	
Hispanic (of any race)	7.0	16.8**
Non-Hispanic white alone	80.1	65.4**
Black Alone	10.6	14.6**
Some Other Race Alone	2.0	5.1**
Two or More Races	0.7	2.4**
Total	100.4	104.3**
Party Identification	$N = 152$	$N = 2{,}805$
Democratic	36.2	33
Independent	30.3	41
Republican	29.6	24
Other	3.9	2
Total	100.0	100
2000 Vote	$N = 151$	Survey was conducted before the election.
Democratic	34.4	
Republican	35.1	
Other	6.0	
Didn't vote	24.5	
Total	100.0	

*Deviations from $N = 153$ are due to nonresponse.
**Data from the 2000 Census for the total population of Florida. Total adds to more than 100% because the categories are nonoverlapping.

certain redundancy that protected against operational problems. For example, if one coder fell ill and was unable to complete the assignment, there would still be two records of the markings to measure both vote totals and reliability.

We trained the coders to work independently of one another and to record what they saw in the form of a numerical code (e.g., code 1 signifies a chad with a single corner detached) for each chad or candidate position on each ballot. For example, with three coders, C chad or candidate positions, and B ballots, there were $3CB$ codes recorded. Experts from NORC and the Media Group created the numeric coding scheme to describe the types of marks on ballots.

The coders were Florida residents who worked in the counties in which they lived, and possibly in adjoining counties. They were drawn from past and present NORC field staff and new employees hired specifically for this assignment. Each candidate for employment was asked to respond to a political activities screener (to ensure that no individuals with undue political bias were hired as coders) and to pass a near-point vision test (to ensure that his or her near-point vision was sufficient for ballot examination). Recruits were also informed that the work would require many hours of sitting and focusing on a very small object held at arm's length. Those who did not pass the political screener or the near-point vision test were not hired. Those

who did not wish to work under the anticipated physical conditions passed on this assignment. Those selected were assigned to teams of three.

We assigned senior-level staff to act as team leaders and to provide on-site supervision for the coding operation. Thus, each team consisted of four people: one team leader who was responsible for supervision and management issues and three coders who were responsible for coding the marks on the uncertified ballots.

We developed and implemented an integrated data capture system that involved paper coding forms and a computerized data-entry system. We designed three coding forms, one for each of the three main voting systems used in Florida (Votomatic, Datavote, and optical scan), to capture chad-level or candidate-level information from the presidential and U.S. Senate sections of the ballots. The forms also allowed us to capture some information from remaining sections of the ballot (e.g., whether there were dimples in other races on the ballot, whether the voter used colored ink, whether the ballot was torn, and verbatim transcription of any notes written on the ballot by the voter).

We trained the team leaders on the technical and administrative aspects of the project over a two-day period. Subsequently, team leaders were responsible for training and evaluating their teams of coders. Team leaders and coders participated in mock coding sessions prior to the start of work. Workers who did not demonstrate acceptable coding in these sessions were replaced.

The intention of the project was not to mimic or replicate the work of county canvassing boards. Our coders examined each chad or candidate position independently of all other chad and candidates and made a judgment about the chad being observed. Conversely, canvassing board members examine a ballot as a committee and discuss the markings. They examine all markings on a ballot in relation to each other with the goal of assigning a vote to a candidate. NORC coders did not attribute votes to candidates.

The ballot examination period extended from February 5, 2001, to May 30, 2001. Between one and four teams worked in parallel in the each of the counties, with four used in the largest counties and one in the smallest. Prior to the arrival of the teams, representatives from the Media Group and local election officials worked to prepare the ballots for examination. In general, the uncertified ballots were organized by ballot type and precinct. Thus, coding teams were able to view all undervotes (by precinct) and then all overvotes (by precinct) in an efficient manner.

Florida law stipulates that only county elections officials may handle ballots. Thus, coders sat on one side of a table and examined ballots that were displayed by officials sitting on the other side of the table. The coders used light boxes to examine dimples and other nuances on the ballots and they instructed officials to display the ballots over the light at advantageous viewing angles.

Team leaders monitored the work of the coders to assure that the forms were being completed as required and that the coders demonstrated careful recording of identifying information. Yet team leaders did not check the accuracy of the coding—rather, the aim was to record accurately the independent judgments of the coders.

During planning stages for the project, experts from the Media Group theorized that overvote ballots would be easier to code with less variability than undervote ballots. They reasoned that full punches and fully filled ovals are typical of overvote ballots and should be very easy to identify. We tested their theory in Polk, Nassau, and Pasco counties—which employed optical scan, Datavote, and Votomatic systems—and found that the three coders agreed in the codes assigned almost always (see Section 4.2 and Table 8 for details). Because the reliability of the overvote ballots was apparently so high, the experts decided there was little benefit to using three codings in all counties. Subsequently, as a cost-saving measure, we used only one coder for overvote ballots in the remaining 64 counties.

At the end of each day, team leaders conducted a second review of the day's work to ensure that all forms were properly completed and organized for transmittal to NORC's data entry facilities in Chicago. Team leaders shipped the forms to Chicago via Federal Express, including a special transmittal form detailing the county worked, the number of forms included, and the identification numbers associated with the shipment.

In Chicago, we entered identifying information about the coding forms into an electronic tracking system and prepared the forms for data entry. Trained clerks reviewed every cell on every form for completeness and legibility. They assigned special codes, if necessary, to missing data or illegible fields.

Trained operators key-entered the coding forms, followed by 100% verification (i.e., all forms were key entered independently a second time). Managers conducted an adjudication/reconciliation process to resolve any discrepancies.

We developed two primary databases for the ballot-level information. One database contains the coded information for every chad or candidate space on every ballot across 67 counties. This file does not attempt to reconcile candidate information across ballots, it simply reflects the reality of the disparate ballot designs used throughout the state of Florida. (By this, we mean that our raw data reflects the fact that the candidates names are not in the same field on each ballot for each county.) The second or aligned database reconciles the coded information for every ballot for each presidential and U.S. Senate candidate. (This means each coder/candidate is in the same field in each ballot record.)

3. CHARACTERISTICS OF THE FLORIDA CODERS

Overall, 153 coders worked on the project. Once hired, each coder completed a short demographic questionnaire. For comparison, we can use the Year 2000 General Social Survey, a national sampling of householders 18 years of age and over. Table 2 shows the results.

When compared with the general population, the coders are more often female and older and have more years of formal education. When one turns to variables with potential political or ideological implications, there are small differences in family income, ethnicity, and party identification. In terms of presidential vote in 2000, coders split right down the middle.

It would be far fetched to claim the coders are a representative sample. To begin with, a truly representative sample of workers would not be available for such temporary employment. But Table 2 supports two propositions: (1) the coders as a group seem to have no political tilt which might influence their collective judgments and (2) the results here seem roughly applicable to

Table 3. Number of Undervote Ballots by Mark Status by Type of Ballot

Ballots by mark status	Votomatic ballots	Datavote ballots	Optical ballots	Total ballots
Total Ballots	53,193	765	7,202	61,150
Completely Blank Ballots *	28,970	466	4,599	34,025
Marked Ballots **	24,223	299	2,603	27,125
Marked Ballots by Candidate #				
Marked for Bush	11,735	168	1,210	13,113
Marked for Gore	11,726	135	1,422	13,283
Marked for Browne	599	8	141	748
Marked for Nader	617	4	164	785
Marked for Harris	347	2	122	471
Marked for Hagel	479	1	107	587
Marked for Buchannan	572	2	125	699
Marked for McReynolds	221	2	107	330
Marked for Phillips	359	4	102	465
Marked for Moorehead	285	2	102	389

* All coders saw no marks for any presidential candidate.
** At least one coder saw a mark for at least one presidential candidate.
# At least one coder saw a mark for the named candidate. Because some ballots were marked for more than one candidate and some were judged marked for different candidates by different coders, marked ballots by candidate do not add to total marked ballots.

the people who become election officials at the precinct level, provided they have training and supervision similar to that of the project and are not political partisans.

4. DESCRIPTION OF CODER RELIABILITY

The objective of the coding of the uncertified ballots was to obtain an accurate (reliable and valid) record of the marks on the ballots. Knowing that no coding operation can ever be flawless, our objective was to assess the quality of the coding operation. This quality has two principal dimensions: reliability and validity.

Reliability is a measure of the consistency in the data; it can be described using a variety of measures, some of which we discuss in what follows. Bailar and Tepping (1972) and Kalton and Stowell (1979) discussed coder reliability studies. The first dimension, *validity* (average correctness), is much more difficult to assess, and there is little opportunity to measure validity from internal evidence within our data. We attempted to avoid invalid data by not hiring biased coders and by coder training and supervision.

Sections 4.1 and 4.2 discuss reliability issues for undervote and overvote ballots, respectively. Throughout these sections, we examine reliability at the candidate/ballot or chad level.

4.1 Reliability of Undervote Ballots

Table 3 shows the number of undervote ballots by mark status and by type (Votomatic, Datavote, and Optical Scan). We use the term *mark status* to signify whether a ballot was completely blank on the presidential race (as determined by all coders who reviewed the ballot) or not. Almost 56% of total ballots were completely blank. The table also shows the number of undervote ballots that were marked (by at least one coder) for each presidential candidate. For example, 13,113 ballots were marked for Bush and 13,283 were marked for Gore.

Blank ballots are uninteresting and uninformative regarding the reliability of Florida's ballot systems. Thus, in the balance of this section, we present statistics only for the universe of 27,125 marked ballots.

For a given universe of ballots, we gauge the reliability of alternative ballot systems using four statistics defined as follows:

$$\overline{P}_2 = \sum_j^J \left(B_{jj+} + B_{j+j} + B_{+jj}\right)/3B,$$

$$\kappa_2 = \frac{\overline{P}_2 - \sum_j^J P_j^2}{1 - \sum_j^J P_j^2},$$

$$\overline{P}_3 = \sum_j^J B_{jjj}/B,$$

and

$$\kappa_3 = \frac{\overline{P}_3 - \sum_j^J P_j^3}{1 - \sum_j^J P_j^3},$$

where B is the total number of ballots considered in the universe; j indexes the code (or variable) values; J is the total number of discreet code values; B_{ijk} is the number of ballots assigned the ith code value by the first coder, the jth value by the second coder, and the kth value by the third coder; and $P_j = (B_{j++} + B_{+j+} + B_{++j})/3B$ is the (assumed common) probability of assigning the jth value. See Fleiss (1965, 1971).

$\overline{P}_2$ is the proportion of agreement among all pairwise comparisons between coders. If three coders examine a given ballot, then there are three pairwise comparisons for each of the 10 presidential candidates, or 30 pairwise comparisons for the ballot overall. Similarly, $\overline{P}_3$ is the proportion of agreement among

Table 4. Reliability Statistics by Candidate by Outcome Variable: Total Marked Ballots

Candidate	All codes				Dimple or greater				Two corners or greater			
	$\overline{P}_2$	κ_2	$\overline{P}_3$	κ_3	$\overline{P}_2$	κ_2	$\overline{P}_3$	κ_3	$\overline{P}_2$	κ_2	$\overline{P}_3$	κ_3
Bush	0.81	0.68	0.73	0.65	0.89	0.77	0.84	0.77	0.98	0.85	0.97	0.85
Gore	0.80	0.65	0.71	0.63	0.88	0.73	0.81	0.73	0.99	0.81	0.98	0.81
Browne	0.98	0.55	0.98	0.58	0.99	0.47	0.99	0.47	1.00	0.48	1.00	0.48
Nader	0.99	0.65	0.98	0.67	0.99	0.62	0.99	0.62	1.00	0.63	1.00	0.63
Harris	0.99	0.60	0.99	0.62	0.99	0.51	0.99	0.51	1.00	0.51	1.00	0.51
Hagel	0.99	0.62	0.98	0.65	0.99	0.55	0.99	0.55	1.00	0.56	1.00	0.56
Buchannan	0.99	0.63	0.98	0.65	0.99	0.64	0.99	0.64	1.00	0.58	1.00	0.58
McReynolds	0.99	0.59	0.99	0.62	1.00	0.46	0.99	0.46	1.00	0.43	1.00	0.43
Phillips	0.99	0.62	0.99	0.64	1.00	0.43	0.99	0.43	1.00	0.49	1.00	0.49
Moorehead	0.99	0.61	0.99	0.63	1.00	0.51	0.99	0.51	1.00	0.47	1.00	0.47

all three-way comparisons. If three coders examine the given ballot, then there are 10 such comparisons, one per candidate. If there are B such ballots, then there are $30B$ pairwise comparisons and $10B$ three-way comparisons over all ballots. Both $\overline{P}_2$ and $\overline{P}_3$ range from 0 to 1, with larger values signifying greater consistency in coding.

κ_2 and κ_3 convey similar, though not identical, information. κ_2 measures the excess of actual pairwise agreement over the agreement expected given a randomization model. The excess is normalized by the maximum possible excess over random agreement. κ_3 is similarly defined but in terms of three-way agreement. Both statistics range from 0 to 1. Small values signify that actual agreement is little above random agreement, while larger values signify that actual agreement is greater than random agreement.

Throughout the balance of this section, we present various reliability statistics and, where possible, draw conclusions about the reliability of different types of undervote ballots. We emphasize that all statements about reliability refer to the process of hand counting (or coding ballots), given that a machine has already classified the ballots as undervotes. Overall, there are two dimensions of reliability that must be of concern to election administration: (1) the extent to which ballots are classified as undervotes, and (2) the consistency of coding ballots, given they are classified as undervotes. The former may be assessed by the undervote error rate, reviewed earlier in Section 1. This section deals strictly with the latter dimension.

Table 5. Reliability Statistics by Candidate by Outcome Variable by Type of Ballot: Total Marked Ballots

Candidate	Dimple or greater		Two corners or greater	
	$\overline{P}_2$	$\overline{P}_3$	$\overline{P}_2$	$\overline{P}_3$
A. Votomatic ballots				
Bush	0.88	0.83	0.98	0.98
Gore	0.87	0.80	0.99	0.98
Browne	0.99	0.99	1.00	1.00
B. Datavote ballots				
Bush	0.79	0.70	0.87	0.81
Gore	0.86	0.79	0.92	0.88
Browne	0.99	0.98	1.00	1.00
C. Optical scan ballots				
Bush	0.96	0.94	0.98	0.97
Gore	0.94	0.91	0.98	0.97
Browne	0.99	0.98	1.00	0.99

Table 4 gives the four reliability statistics for each of the ten named presidential candidates and for each of three outcome variables. The first variable—*all codes*—refers to the original coding by the coders. For this variable, agreement is measured at the level of the raw code value. The second and third variables are derived from this first variable. *Dimple or greater* is an indicator variable defined equal to 1, if the original code signifies a dimple or greater, and to 0 otherwise, while *two corners or greater* is defined equal to 1, if the original code signifies two corners detached or greater, and to 0 otherwise. For these variables, agreement is measured in terms of the recoded values, not in terms of raw code values. Dimple or greater corresponds to a voting standard advocated by the Gore team during the presidential election, while two corners or greater corresponds to a voting standard advocated by the Bush team (see, e.g., *Orlando Sentinel* or *Wall Street Journal*, both of November 12, 2001).

As expected, we find that agreement is at a relatively lower level for the all codes variable than for the dimple or greater variable than for the two-corners or greater variable. The all codes variable provides the most detailed description of a candidate/ballot, and agreement for this variable must be a relatively rarer event than agreement for either of the other two variables. Dimple or greater is possibly less obvious and relies more on coder judgment than two corners or greater. The concepts of dimpling and two-corner detachment apply most directly to punch card systems, including Votomatic and Datavote ballots. They do not have exact counterparts for optical scan ballots. To continue our analyses, we defined these outcome variables as any affirmative mark and a completely filled arrow or oval, respectively. We believe these definitions make the outcome variables reasonably, though not exactly, comparable across the three types of ballots.

Reliability is relatively similar for Bush and for Gore. For example, $\overline{P}_2$ is 0.89 and 0.88 and κ_2 is 0.77 and 0.73 for Bush and Gore, respectively, given the dimple or greater variable. Reliability is apparently at a much higher level for each of the remaining eight candidates. For example, under the dimple or greater standard, Browne's $\overline{P}_2$ is 0.99. This apparent higher reliability is due to the fact that few voters intended to vote for these candidates and most of the variable values for these candidates are 0.

Table 6. Reliability Statistics by Candidate by Outcome Variable by Absentee Status: Total Marked Ballots

Candidate	Dimple or greater $\overline{P}_2$	Dimple or greater $\overline{P}_3$	Two corners or greater $\overline{P}_2$	Two corners or greater $\overline{P}_3$
A. Total marked ballots				
A.1. Absentee ballots				
Bush	0.86	0.79	0.97	0.96
Gore	0.87	0.80	0.98	0.97
Browne	0.99	0.99	1.00	1.00
A.2. Non-absentee (regular) ballots				
Bush	0.90	0.85	0.99	0.98
Gore	0.88	0.82	0.99	0.98
Browne	0.99	0.98	1.00	1.00
B. Votomatic marked ballots				
B.1. Absentee ballots				
Bush	0.85	0.78	0.98	0.97
Gore	0.86	0.78	0.99	0.98
Browne	0.99	0.99	1.00	1.00
B.2. Non-absentee (regular) ballots				
Bush	0.89	0.84	0.99	0.98
Gore	0.87	0.81	0.99	0.98
Browne	0.99	0.98	1.00	1.00
C. Datavote marked ballots				
C.1. Absentee ballots				
Bush	0.77	0.66	0.86	0.79
Gore	0.85	0.77	0.92	0.88
Browne	0.99	0.98	1.00	1.00
C.2. Non-absentee (regular) ballots				
Bush	0.98	0.97	0.98	0.97
Gore	0.94	0.92	0.94	0.92
Browne	1.00	1.00	1.00	1.00
D. Optical scan marked ballots				
D.1. Absentee ballots				
Bush	0.97	0.95	0.98	0.97
Gore	0.97	0.95	0.97	0.96
Browne	0.99	0.99	1.00	0.99
D.2. Non-absentee (regular) ballots				
Bush	0.96	0.93	0.98	0.97
Gore	0.93	0.90	0.98	0.97
Browne	0.99	0.98	1.00	1.00

In what follows, we only present reliability statistics for the Republican, Democratic, and Libertarian candidates: Bush, Gore, and Browne. Results for other candidates mirror those for Browne. To simplify the presentation, we present reliability statistics only for the dimple or greater and two corners or greater variables. Also, because of the similarities between them, we drop the κ statistics and only present the $\overline{P}$ statistics.

Reliability by type of ballot appears in Table 5. For the dimple or greater standard, the reliability of the two punch card ballots is lower than that of the optical scan ballots. For the two-corners or greater variable, the reliability for Datavote (punch card without chad) is lower than for the other two types of ballots. Indeed, the two variables are quite similar for Datavote ballots, tending to differ only for absentee ballots.

While Bush and Gore reliabilities are similar for Votomatic and optical scan ballots, Bush reliability is lower than Gore reliability for Datavote ballots. This finding is probably due to the fact, that only 299 Datavote ballots were marked at all; that 135 were marked for Gore; and that 168 were marked for Bush. Thus, more ballots were blank for Gore—on which there was complete agreement between the coders—leading to relatively higher Gore reliability.

Table 6 presents reliability statistics by the absentee status (absentee versus regular) of the ballots. Overall, absentee ballots are slightly less reliable than regular ballots. Given the dimple or greater standard, the Bush and Gore pairwise agreement statistics are 0.86 and 0.87 for absentee ballots and 0.90 and 0.88 for regular ballots, respectively. All four reliability measures reveal the slightly greater reliability of regular ballots. For the two-corners or greater standard, the reliability of absentee and regular ballots tend to converge.

Votomatic ballots tend to follow the pattern by absentee status observed overall. Datavote ballots, however, reveal larger differences, but still regular ballots are the more reliable. For example, given the dimple or greater standard, the Bush pairwise agreement statistics are 0.77 for absentee ballots and 0.98 for regular ballots, respectively. The differences even continue for the two corners or greater standard. Datavote absentee ballots also reveal a relatively sizable difference between Bush and Gore reliability. As noted earlier, this finding is probably due to the fact that more of these ballots were blank for Gore. Optical scan ballots reveal negligible differences between absentee and regular ballots.

Table 7 examines various design formats used in optical scan counties. Some counties used an arrow design, while others used an oval design. From these data, it appears the oval design is slightly the more reliable. For example, given the dimple or greater standard, the Gore and Bush pairwise agreement statistics are 0.95 and 0.92 for the arrow design and 0.97 and 0.95 for the oval design, respectively. As we have seen before, observed

Table 7. Reliability Statistics by Candidate by Outcome Variable by Design Format: Optical Scan Marked Ballots

Candidate	Dimple or greater $\overline{P}_2$	Dimple or greater $\overline{P}_3$	Two corners or greater $\overline{P}_2$	Two corners or greater $\overline{P}_3$
A. Arrow design				
Bush	0.95	0.92	0.98	0.97
Gore	0.92	0.88	0.99	0.98
Browne	0.99	0.98	0.99	0.99
B. Oval design				
Bush	0.97	0.95	0.98	0.98
Gore	0.95	0.93	0.97	0.96
Browne	0.99	0.99	1.00	1.00
C. Oval design, single column				
Bush	0.96	0.94	0.98	0.97
Gore	0.95	0.93	0.97	0.96
Browne	0.99	0.98	1.00	1.00
D. Oval design, split column				
Bush	0.97	0.96	0.99	0.98
Gore	0.94	0.91	0.96	0.95
Browne	0.99	0.99	1.00	1.00

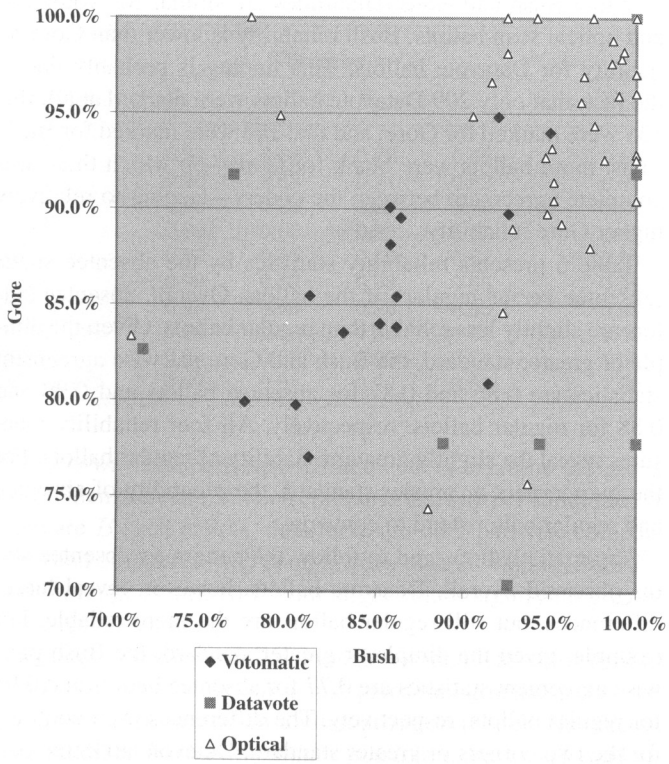

Figure 1. Scatterplot of Florida Counties, Gore $\overline{P}_2$ versus Bush $\overline{P}_2$, Dimple or Greater Standard: Total Marked Ballots.

differences for the dimple or greater standard tend to converge for the two-corner or greater standard.

About three-fourths of the oval-design ballots listed the presidential candidates in a single column, and about one-quarter split the candidates across two columns. One might expect the single column design to be the more reliable, yet the statistics reveal that both of these variations on the oval design are about equally reliable.

In Figure 1, we study reliability by county, plotting Gore pairwise agreement versus Bush pairwise agreement, given the dimple or greater standard. We have obtained similar results for the other reliability statistics and for the two corners or greater standard. Four counties—Calhoun, Dixie, Hendry, and Wakulla—did not have any marked ballots, and thus they are omitted from the figure. For Bush, agreement varies from just over 0.70 in Baker county to essentially 1.00 in Union and other counties. For Gore, agreement varies from just over, 0.70 in Jefferson county to essentially 1.00 in Union and other counties.

If Gore and Bush agreements perfectly tracked one another (high for Bush also high for Gore and low for Bush also low for Gore) then the various points, each of which represents a county, would fall on a straight line. One can see reasonably close tracking for the Votomatic counties: the counties (or points) do tend to fall along a straight line. The Datavote counties are generally quite small, and the pairwise agreement statistics tend to be a bit erratic because they are based upon a relatively small number of ballots. The optical scan counties form a large cluster with very high agreement for both Bush and Gore, and a smaller number of counties with differential agreement.

The analysis by county is limited by the fact that, in general, different teams of coders worked in different counties. Because there was no explicit random assignment of coder-teams to counties, county effects and coder-team effects are confounded with one another.

4.2 Reliability of Overvote Ballots

Finally, we turn to a brief examination of overvote ballots. Recall that we decided to use three coders for overvotes in only three of Florida's counties, and one coder per ballot in the remaining 64 counties. They examined 2,114 Votomatic ballots in Pasco County, 1,294 Datavote ballots in Nassau County, and 668 optical scan ballots in Polk County, or 4,076 ballots overall. All reliability statistics are very high, regardless of candidate, ballot type, or outcome variable. From Table 8 we see that pairwise agreement is about 0.97, 1.00, and 0.99 for Votomatic, Datavote, and optical scan ballots, respectively. Apparently, all ballot systems have similarly high reliability given ballots have been classified as overvotes.

5. EFFECTS OF CODER CHARACTERISTICS ON CODER RELIABILITY

The coding of the data and the supervision of the coding were carried out with care and with the purpose of preventing any deliberate miscoding of ballots. However, there is always a concern that some characteristic of the coders may be associated systematically with some failure of accurate coding; for example, older coders might have more difficulty with their eyesight, and might miss more non-null codes than their younger co-workers. Of greater concern would be the possibility of systematic favoritism toward a particular candidate, either deliberate or subconscious. Though coders did not have access to the relationship between chads and candidates in Votomatic counties, this information would not be difficult to obtain, and could in any case become known to coders in an innocent manner.

Though NORC collected and recorded "extra-role" characteristics of coders—that is, characteristics of the coders that are not related to their role as coders [see, e.g., Sudman and Bradburn

Table 8. Reliability Statistics by Candidate by Outcome Variable by Type of Ballot: Overvote Ballots

	Dimple or greater		*Two corners or greater*	
Candidate	$\overline{P}_2$	$\overline{P}_3$	$\overline{P}_2$	$\overline{P}_3$
	A. Three-county total			
Bush	0.98	0.97	0.98	0.97
Gore	0.98	0.98	0.98	0.98
Browne	0.98	0.97	0.98	0.97
	B. Votomatic ballots (Pasco County)			
Bush	0.97	0.96	0.97	0.96
Gore	0.97	0.96	0.98	0.96
Browne	0.97	0.96	0.97	0.96
	C. Datavote ballots (Nassau County)			
Bush	1.00	1.00	1.00	1.00
Gore	1.00	1.00	1.00	1.00
Browne	1.00	0.99	1.00	1.00
	D. Optical scan ballots (Polk County)			
Bush	0.99	0.98	0.99	0.98
Gore	0.99	0.98	0.99	0.98
Browne	0.98	0.97	0.98	0.97

(1974) for analogous characteristics of survey interviewers]—they were not used in any way to allocate work to coders. Thus, in some counties and precincts all three coders might have been similar in age and socio-economic characteristics. In other counties and precincts the three coders might have differed on all the characteristics. In still others they might have differed on some characteristics and not on others.

This nonsystematic work allocation means that it is very difficult to establish the possible effect of extra-role characteristics of coders. It is not sufficient simply to compare the overall rate of codings for one candidate or another across all the ballots. Ballots differ from one another not only in level of difficulty (e.g., by voting system, quality of voting machines, maintenance of equipment, and haphazard variation in level of difficulty of coding) but also systematically by the voting propensities in the precinct to which they belong.

It is therefore necessary to construct a model to estimate the potential impact of coder characteristics. What is needed is a model that estimates the effect of each characteristic taking into account (controlling for) all the effects of the other characteristics. Thus, if we wish to estimate the effect of age of coder, for example, we would like to contrast the codings of younger and older coders while coding equivalent sets of ballots. The analysis would also need to control for all the other characteristics simultaneously; thus only young and old coders with otherwise identical characteristics (gender, socio-economics, party affiliation) could be used in the comparison. To carry out the contrast only in such equivalent sets would mean discarding most of the data.

Statistical analysis generally deals with this class of problem by using some form of regression analysis—fitting a model that accounts simultaneously for a set of characteristics. There are two particular issues that complicate the analysis for these data.

First, the outcome variable (the code) is an attribute, a binary variable that takes only two values, yes (there is a mark that has been coded) and no (there is no mark that has been coded). This outcome variable implies the use of logistic regression. Second, the data have a hierarchical structure that arises in two conceptually different ways:

1. Ballots are grouped within precinct; precincts are natural groupings of ballots that have their own (mostly unknown) characteristics. Precincts are grouped within counties; counties are also natural groupings with potentially distinctive features. Groups of counties share voting technology.

2. The three codings of the ballots are repeated measurements on the set of underlying ballots. Furthermore, teams of coders operate within precinct (coders are nested within precinct); generally, the same team of coders coded all the ballots in a precinct.

The hierarchical structure of the data implies violations of some of the usual assumptions of linear regression that are needed to obtain good estimates of effects, and in particular, good estimates of the precision of the estimators. The overall model is therefore a repeated measures multilevel model. These models were explored in depth by Goldstein, Healy, and Rasbash (1994) and Goldstein (1995).

The model we fit is described in the following. At each level of the hierarchy we allow for variation in the intercept of the regression equation. The slope is invariant across levels of the model. Thus, we estimate a single regression coefficient for the impact of each of the explanatory variables in the model.

The outcome variable for a given candidate is defined by

$Y_{ijkp} = 1$, if the ith coder in the jth precinct in the kth county sees a mark for the candidate on the pth ballot;
$= 0$ otherwise.

The outcome variable is modeled as a Bernoulli random variable with parameter $\phi_{ijkp} = \Pr\{$the ith coder, jth precinct, kth county sees a mark for the candidate on the pth ballot$\}$.

Define the log-odds by $\eta_{ijkp} = \log\{\phi_{ijkp}/(1-\phi_{ijkp})\}$, and define a hierarchical, generalized linear model for the log-odds as follows:

Level 1: Coder Level

$$\eta_{ijkp} = \beta_{0jkp} + \sum_q \beta_{qjkp} X_{qijkp}, \tag{1}$$

where $q = 1, 2, \ldots, Q$ indexes the coder-level characteristics denoted by X;

Level 2: Ballot Level

$$\beta_{0jkp} = \lambda_{0jk0} + w_{0jkp}, \tag{2a}$$

$$\beta_{qjkp} = \lambda_{qjk0}, \tag{2b}$$

where $w_{0jkp} \sim N(0, \kappa)$ and κ is the between-ballot variability within the same precinct;

Level 3: Precinct Level

$$\lambda_{0jk0} = \alpha_{00k0} + u_{0jk0} \tag{3a}$$

$$\lambda_{qjk0} = \alpha_{q0k0}, \tag{3b}$$

where $u_{0jk0} \sim N(0, \tau)$ and τ is the between-precinct variability within the same county; and

Level 4: County Level

$$\alpha_{00k0} = \gamma_{0000} + \nu_{00k0} \tag{4a}$$

$$\alpha_{q0k0} = \gamma_{q000}, \tag{4b}$$

where $\nu_{00k0} \sim N(0, \xi)$ and ξ is the between-county variability of coder-level effects.

In this notation, the subscript value "0" in a certain subscript position signifies that the parameter in question is fixed over entities represented by that position. For example, β_{0jkp} varies over precincts, counties and ballots, but not over coders, and α_{00k0} varies only over counties.

The following coder characteristics were included in our analyses as X variables: gender (binary: female as reference category); marital status (binary: not married as reference category); age (in years); education (binary: not a college graduate as reference category); income (binary: income $<$ \$50,000 as reference category); race (binary: nonwhite as reference category); political affiliation (three categories: Democrat as reference category; Republican; neither Democrat nor Republican).

Table 9. Multilevel Repeated Measures (Four-Level: Coder-Ballot-Precinct-County) Regression in Votomatic Counties: Coefficients (Standard Errors)

Outcome Variable	Gender	Marital status	Age	Education	Income	Race	Political affiliation: Republican	Political affiliation: Neither
Bush	0.335* (0.019)	0.250* (0.019)	0.002* (0.001)	−0.026 (0.018)	−0.211* (0.021)	−0.186* (0.028)	0.220* (0.019)	0.128* (0.021)
Gore	0.066* (0.018)	−0.010 (0.0184)	−0.002* (0.001)	−0.046* (0.017)	0.016 (0.020)	−0.041 (0.026)	−0.040* (0.018)	0.014 (0.019)
Difference	0.031* (0.004)	0.027* (0.004)	0.000 (0.000)	0.005 (0.004)	−0.026* (0.005)	−0.008 (0.006)	0.031* (0.004)	0.012* (0.005)

*Coefficient is statistically significant at the 0.05 level.

The basic purpose of the models is to partition the variability in the data so that we can estimate the impact of each coder characteristic while simultaneously taking into account both the differences between precincts and counties and the clustering of ballots within precincts and counties. The full model is a three-level model with repeated measures at the first level, which is in effect a four-level model for estimation purposes. We used the current (2001) version of the MLWin software (Goldstein et al. 1998) to fit the models.

If we were to fit a model to all the Florida data simultaneously we would need a five-level model: repeated measures (1), within ballot (2), within precinct (3), within county (4), within voting technology (5). The five-level model presented technical problems in estimation, and there seemed insufficient advantage in combining different voting technologies in a single analysis. As it happens, the results and the implications are different for the different technologies, and thus it is best in any case to look separately at the results. There were insufficient cases in the Datavote counties for the models to run successfully; consequently, we report separately for only Votomatic and optical scan ballot systems.

As outcome variables we considered "any mark" for Bush; "any mark" for Gore; and the difference between the coding for Bush and the coding for Gore. We should point out that the final dependent variable (the difference) is not a binary variable and we used a linear rather than a logistic regression in estimating this model.

In each case we ran (1) single-level (nonhierarchical) models, ignoring the hierarchical structure; (2) two-level models, taking into account only the repeated measures aspect of the data; (3) three-level models, incorporating the precinct clustering into the analysis; and (4) four-level models, taking into account the clustering of precincts into counties. We carried out the analyses separately for optical scan counties as a group and Votomatic counties as a group.

For optical scan counties, the one-level (no structure) and two-level (repeated measurement structure) models showed significant effects of coder characteristics on the outcome of the coding operations. If we were to accept these results as valid, we would conclude that coder characteristics do have a significant impact on the results of the coding process. In particular we would believe that the coding is sensitive to the extra-role characteristics of the coders. However, once we included the

Table 10. Odds Ratios (with Confidence Intervals) and Adjusted Probabilities of Coding by Coders of Each Party Affiliation for Votomatic Undervote Ballots

Outcome variable/ political affiliation of the coder		Odds ratio	95% confidence interval	Adjusted probability
Bush/	Republican	1.247	(1.201, 1.295)	0.192 = Pr (Mark for Bush | Coder is Affiliated with Republican Party)
	Neither	1.137	(1.092, 1.184)	0.178 = Pr (Mark for Bush | Coder is Affiliated with neither party)
				0.160 = Pr (Mark for Bush | Coder is Affiliated with Democratic Party)
Gore/	Republican	0.961	(0.927, 0.996)	0.182 = Pr (Mark for Gore | Coder is Affiliated with Republican Party)
	Neither	1.014	(0.976, 1.053)	0.190 = Pr (Mark for Gore | Coder is Affiliated with neither party)
				0.188 = Pr (Mark for Gore | Coder is Affiliated with Democratic Party)

Table 11. Vote Margins for Official Results and Nine Recount Scenarios

Scenario	Description	Margin, Majority Agreement	Margin, Unanimous Agreement
Official results	Vote totals certified by Florida election officials	Bush: +537	
3. Florida Supreme Court Complex	Accepts completed recounts in eight counties, certified results from four counties that said they would not have recounted, and applies 55 county-reported standards; includes overvotes where counties report intent to count them. If the U.S. Supreme Court had upheld the Florida Supreme Court ruling, this scenario would have resulted.	Bush: +493	Bush: +323
2. Florida Supreme Court Simple	Accepts completed recounts in four counties; elsewhere punch card undervotes with at least one corner detached; any affirmative mark on optical scan ballots; no overvotes	Bush: +430	Bush: +369
8. Gore Request	Accepts certified results in 65 counties, including hand-counts in Broward and Volusia; uncertified hand-counts in Palm Beach and parts of Miami-Dade; and one-corner detachment in remaining Miami-Dade; no overvotes	Bush: +225	Bush: +212
4. 65 County Custom	Accepts whatever each individual county considered a vote, includes overvotes where appropriate; certified results accepted for Broward and Volusia	Gore: +171	Gore: +81
7. Most restrictive	Requires unequivocal punches and complete fills on optical scan ballots; no overvotes; accepts certified results Volusia	Gore: +115	Gore: +127
6. Most Inclusive	Accepts all dimpled chads; any affirmative mark on optical scan ballots; includes overvotes in optical counties if prevailing standard dictates; accepts certified results for Volusia	Gore: +107	Bush: +110
5. Bush Standard	Accepts chads with two corners detached; any affirmative mark on optical scan ballots, includes overvotes; accepts certified results for Volusia	Gore: +105	Gore: +146
1. Prevailing Statewide	Requires one corner detached on punch cards; any affirmative mark on optical scan ballots; includes overvotes; accepts certified results for Volusia	Gore: +60	Gore: +145
9. Palm Beach Rules	Accepts dimpled chads only when other dimples present on ballot; includes overvotes in optical scan counties if prevailing standard dictates; accepts certified results for Volusia	Gore: +42	Gore: +114

Source: *Chicago Tribune*, November 12, 2001.

hierarchical structure in the model, all the significant effects disappeared. Thus there is no evidence of systematic effects of coder characteristics on the outcome of the coding process in optical counties.

For Votomatic counties, the one-level (no structure) and two-level (repeated measurement structure) models again showed significant effects of coder characteristics on the outcome of the coding operations. For those counties, however, even when we introduced the hierarchical structure, the effects did not disappear. The coefficients and their standard errors are shown in Table 9. For example, the log-odds of seeing a mark for Bush is significantly higher for male coders than for female coders. The log-odds of recording a mark for Gore is also higher for males than for females. But the sex effect is smaller in the Gore model than in the Bush model.

For the logistic regression models, the dependent variable is the log of the ratio of the proportion coding for a particular candidate to the proportion not coding for that candidate; this ratio

Table 12. New Votes Found by the Media Group Among Uncertified Ballots

	Agreement level					
	At least two of three coders			All three coders		
Scenario	Total new votes for Bush and Gore	Percent of certified votes for Bush and Gore	Percent of uncertified ballots	Total new votes for Bush and Gore	Percent of certified votes for Bush and Gore	Percent of uncertified ballots
1	7,811	0.13	4.5	6,542	0.11	3.7
2	5,383	0.09	3.1	4,234	0.07	2.4
3	7,582	0.13	4.3	5,352	0.09	3.1
4	10,480	0.18	6.0	6,810	0.12	3.9
5	7,710	0.13	4.4	6,537	0.11	3.7
6	24,240	0.42	13.9	14,281	0.25	8.2
7	5,332	0.09	3.0	5,094	0.09	2.9
8	1,434	0.02	0.8	1,369	0.02	0.8
9	14,503	0.25	8.3	8,603	0.15	4.9

Source: *Chicago Tribune*, November 12, 2001.

is the odds of coding "yes" for the candidate. The coefficients in the logistic model express the additive impact of each factor on the log of the odds. The underlying model is a multiplicative model for odds; by exponentiating the (additive) coefficient in the logistic regression we can find the multiplicative effect on the odds themselves. For each explanatory variable (factor) there is a reference category. The impact of the factor is expressed as the ratio of the odds when a coder belongs to a particular category relative to the odds when a coder is in the reference category. In particular, for gender the reference category is female; the impact of gender in the model is the ratio of the odds that a male coder will code "yes" to the odds that a female coder will code "yes." This is the odds ratio.

Table 10 considers a detailed breakdown of the impact of political affiliation on the probability of the outcome of the coding operation in the Votomatic counties. (Similar analysis could be done of the gender effect.) The reference category is Democrat—in other words the other two affiliations are contrasted with Democrats in the model. For the Bush model, the coefficient for Republican in the logistic regression model is 0.22. This translates into an odds ratio of 1.25 in Table 10; the 95% confidence interval for this odds ratio is (1.20, 1.29). A null value for the odds ratio would be 1.00 (no effect on the odds).

To understand its substantive significance, we need to translate this odds ratio into an effect on the probability of the coding outcome. This is done in the far right column in of the table. The adjusted probability is the estimated proportion of ballots that will be coded positively (i.e., marked) for the candidate. There are three adjusted probabilities for each outcome. The first is the proportion of ballots that would be coded for the candidate if the coder were Republican. The second is the corresponding proportion if the coder is neither Republican nor Democrat. The third is the proportion if the coder is Democratic. Thus, we can see that in interpreting Bush chads, Republican coders will on average code 19.2% of the ballots as marks for Bush, Democrat coders will code only 16% for Bush; and the others will code 17.8% for Bush. In interpreting Gore chads, the position is reversed, though the size of the effect is much less. Here, 18.2% of Republicans will code a mark for Gore, 18.8% of Democrats will do so, and 19% of the others will code for Gore.

The linear regression model is one way of interpreting the impact of these findings on the difference between for Bush and for Gore. The estimated impact of being Republican rather than Democrat can be seen in the coefficient of 0.03 for Republican in the equation. This implies that on average, Republican coders found 3% more marks for Bush relative to Gore than did Democrat coders.

The results of the multilevel analysis are important in two ways. First, the contrast between the hierarchical and the non-hierarchical analyses show the importance of taking the data structure into account. Failing to take the structure into account could lead to making the wrong inference from the data. Second, there is a clear difference between the outcome in the optical scan counties and the Votomatic counties. In the Votomatic counties there is clear evidence of sensitivity of the coding outcome to the extra-role characteristics of the coders.

6. NEW VOTES

On November 12, 2001, the Media Group published vote totals for Bush and Gore based upon nine possible standards (or scenarios) for what constitutes a vote, with two variations for two of the standards. Their vote totals were defined as the certified votes from the presidential election plus presumed "new votes" discovered among the uncertified ballots, given the various standards. For each scenario, they examined vote totals under two levels of agreement: (1) a unanimous level that required agreement by all three coders, and (2) a majority-rules level that required agreement by at least two of the coders.

Table 11 details the nine scenarios analyzed by the Media Group and corresponding margins of victory. For example, given the Gore Request, Bush's margin is +225 votes. It is important to note that no one scenario was a likely outcome of the actual election. Indeed, a time- and labor-intensive examination and recount of the 175,010 uncertified ballots (such as the one conducted for this project) was not an option in the hectic days following the presidential election. With that caveat in mind, the results of the recount analysis are fascinating and three points merit mention. First, with margins ranging from 42 to 493 votes, no recount scenario produces a margin of victory greater than that of the official certified results. Second, it

appears that each candidate's preferred recount standard would have yielded a defeat for that candidate, for example, Gore wins by 105 votes given the Bush Standard. Third, it is clear that the election was an exceedingly close contest, regardless of which standard would have been used in recounting votes. Readers interested in a more comprehensive or interpretive analysis of the data should see any Media Group member's November 12, 2001 publication and Keating (2002). The online databases from this project allow future researchers to both replicate these findings and investigate alternate scenarios.

Table 12 displays the total number of new votes found by the analysts using NORC's database of uncertified ballots, given each scenario and level of agreement. For example, Bush and Gore collectively got 7,811 new votes according to Scenario 1 and the majority-rules level of agreement, which corresponds to 0.13% of the certified votes cast for these major candidates and 4.5% of all the uncertified ballots.

Leaving aside the substance of the various standards, what is striking to us is that there was apparently such a large volume of new votes among the uncertified ballots. Even under the unanimous level of agreement (which finds fewer new votes than the majority-rules level), there are between 1,369 and 14,281 new votes, which translate into between 0.8% and 8.2% of all uncertified ballots, respectively.

7. SUMMARY

In this article, we obtained the following findings regarding coder reliability. For ballots classified as undervotes:

- Reliability is similar for the two major candidate, Bush and Gore, both statewide and by county, except for some of the smaller counties where small sample size can lead to divergent results.
- Reliability is similar and actually higher for all remaining candidates (than for the major candidates), reflecting the fact that most ballots are blank for these candidates and most coders readily discern and record a blank.
- Reliability is lower for the all codes variable than for the dimple or greater variable, and is highest for the two-corners or greater variable.
- Reliability is lower for the two punch card systems, Votomatic and Datavote, than for the optical scan systems.
- Overall, regular ballots are more reliable than absentee ballots; Datavote absentee ballots are considerably worse than the corresponding regular ballots; absentee and regular ballots are essentially equally reliable for optical scan systems; and Votomatic absentee ballots are slightly less reliable than the corresponding regular ballots.
- Among optical scan ballot designs, the oval is slightly preferred to the arrow.
- Among oval designs, there is little to choose between the single- and split-column formats.

Coder reliability is extremely high for ballots classified as overvotes. Also, undervote and overvote error rates are higher for punch card systems than for optical scan systems. We recommend states consider these findings as they look for ways to reform ballot systems for future elections.

We also found that extra-role characteristics of coders affect coding outcomes in Votomatic, but not optical scan, counties. In Votomatic counties, our analysis shows

- Men see more marks than women.
- Self-identified Republicans are more likely to code a mark in favor of Bush than are self-identified Democrats, while Republicans are less likely to code a mark in favor of Gore than are Democrats.
- Both of the aforementioned findings are stronger in the Bush direction.

This analysis does not tell us which group of coders is correct; it merely identifies systematic differences between groups (or categories) of coders. Neither does it tell us that the characteristics of the coders that we have identified are causally connected to the outcome. It does, however, suggest that the coding of Votomatic ballots is subject to variation that is associated with the characteristics of the coders. This renders adjudication of disputed ballots much more problematic for ballots using the Votomatic system.

Given a number of alternative standards as to what constitutes a vote, analysts working for the Media Group found thousands of potential new votes among the uncertified ballots in NORC's database. These results suggest that state elections officials should consider wider use of new technology that edits the completed ballot in real time as the voter submits it. This technology provides the voter a chance to correct the ballot if it is deemed to be spoiled. Officials should also consider a statewide, consistently applied procedure for human review and counting, where appropriate, of ballots that remain uncertified after the initial machine count.

Arising out of the material here, an important question is the degree to which systems of lower reliability will be tolerated in our democratic system. In this regard, discrepancies between regular versus absentee ballot reliability might provide a benchmark or reference point for future decisions about whether varying ballot methods conform with the constitutional "equal-protection" principal that figured in the ultimate decision by the U.S. Supreme Court.

Finally, in the course of doing this project we observed two other possible limitations in American elections. First, ballots are not usually pretested on real voters prior to their use in the general election. County election officials, some with considerable experience, design ballots under state law. But even experts may create confusing designs or unclear instructions. In the field of survey research, experts draft the questionnaire, pretest it on a small sample, and make appropriate adjustments prior to finalizing it and using it in the main survey. This is parallel to the proof-of-concept stage in technological fields. We recommend the practice of pretesting be extended to the development of ballots prior to their use in a general election. Second, voters are not usually trained prior to an election. Precincts often display voting instructions at the polling place, and precinct workers often

offer assistance to voters who make a specific request. But many people do not take time to read instructions or are embarrassed to ask for help. At the same time, some voters are voting for the first time ever, for the first time in years, or for the first time in a new location under a different ballot system. Some voters' eyesight may be failing. And there can be many other problems too. To address this range of concerns, we recommend states look for cost-effective means of training voters on the ballot system prior to the election.

[Received July 2002. Revised December 2002.]

REFERENCES

Bailar, B. A., and Tepping, B. J. (1972), *Effects of Coders*, Series ER 60 No. 9, Washington, DC: U.S. Bureau of the Census.

Chicago Tribune (2001), "Still Too Close to Call; Conclusion Not Clear Even if Recount Allowed," November 12, pp. 14–15.

Fleiss, J. L. (1965), "Estimating the Accuracy of Dichotomous Judgments," *Psychometrika*, 30, 469.

——— (1971), "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin*, 76, 378.

Goldstein, H. (1995), *Multilevel Statistical Models*, London: Edward Arnold.

Goldstein, H. M., Healy, J. R., and Rasbash, J. (1994), "Multilevel Time Series Models With Applications to Repeated Measures Data," *Statistics in Medicine*, 13, 1643.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., et al. (1998), *A User's Guide to MlwiN*, London: Institute of Education.

Kalton, G., and Stowell, R. (1979), "A Study of Coder Variability," *Applied Statistics*, 28, 276.

Keating, D. (2002), "Democracy Counts: The Media Consortium's Florida Ballot Project," paper prepared for presentation at the annual meeting of the American Political Science Association.

Orlando Sentinel (2001), "The Final Report of the Florida Ballot Project," November 12, p. B5.

Sudman, S., and Bradburn, N. M. (1974), *Response Effects in Surveys*, Chicago: Aldine.

Wall Street Journal (2001), "In Election Review, Bush Wins With No Supreme Court Help; Majority of Florida Voters Would Have Picked Gore but for Poor Ballot Design; Both Backed Wrong Strategy," November 12, p. A14.