Democracy Counts

The Media Consortium Florida Ballot Project

Dan Keating

The Washington Post

keatingd@washpost.com
1150 15th St NW
Washington, DC 20071

Abstract. The contents of ballots that were not counted as a vote for president in Florida's contested 2000 election were catalogued by the National Opinion Research Center under the direction of a consortium of wire service, television and newspaper journalists. Results indicate likely outcomes if the ballots had been recounted under various standards and scenarios. Results also indicate patterns important for election reform and conduct of elections concerning racial differences in voter error, failure rates of different technologies and ballot designs, subjectivity of recounts and validity of mismarked ballots as votes.

Paper prepared for presentation at the annual meeting of the American Political Science Association, Boston, August 2002.

Life is not perfect. Elections are part of life.

Conny McCormack
Election Supervisor
Los Angeles County, CA

The contested Presidential election in Florida in November-December 2000 introduced a new reality to American democracy – the fact that every ballot *doesn't* count. Several technical and administrative obstacles render about 1-in-50 ballots mute.

The controversy over the recounts – a mandatory machine recount, completed and incomplete county recounts, and an aborted statewide recount of selected ballots – focused attention on how voters and voting technology fails.

Even before the Supreme Court ended the disputed election, news organizations had begun to request access to the ballots themselves, a public record under Florida's expansive Sunshine Law (Florida law 101.572).

This paper is intended to introduce people to the datasets gathered that are available for further analysis, and to present analysis conducted by The Washington Post.

Methodology

An media consortium gathered to pool efforts in conducting an examination of the ballots that would be thorough, scientific and transparent – rendering the definitive historic archive of what was on the Florida ballots. The organizations that eventually formed the coalition were the Associated Press, CNN, The Wall Street Journal, The New York Times, The Washington Post, The St. Petersburg Times, The Palm Beach Post and Tribune Publishing, which includes the Los Angeles Times, South Florida Sun-Sentinel, Orlando Sentinel and Chicago Tribune. The group was managed on a consensus basis by a Steering Committee made up John Broder of the New York Times, Doyle McManus of The Los Angeles Times, Bill Hamilton of The Washington Post, Alan Murray and subsequently Phil Kuntz of the Wall Street Journal, Tom Hannon of CNN, Kevin Walsh of Associated Press, Chuck Murphy of the St Petersburg Times and Bill Rose of The Palm Beach Post. Ford Fessenden of the New York Times and Dan Keating of The Washington Post were assigned by the group to design the methodology, obtain staffing and manage the effort on the ground. Murphy helped coordinate with Florida's 67 counties in securing access to the ballots.

Having to devise a unique methodology for collecting this data, the consortium decided to:

- Publicly release all data generated by the effort so that political scientists, political activists and election reformers could replicate any analysis or perform their own review of the ballot contents.
- Review all undervotes and overvotes. Undervotes are ballots on which no presidential choice was detected. Overvotes are ballots on which more than one choice was detected and, thus, no presidential vote was recorded.
- Hire independent researchers to conduct the ballot examination and data collection.
- Blind itself to the data throughout the collection period to prevent leaks. Release the data to members and the public only after it had all been collected.
- Have three people independently review each ballot to measure the subjectivity of recounts and prevent bias in the study. Include the work of all three reviewers in the public data. After ballots were reviewed under each of the major voting

technologies, analysis was done and the methodology was refined to use three reviewers for all undervotes where partial marks presented an issue of discernment, but use one viewer for overvotes where multiple marks were clear enough to be read by machine and were rarely disputed.

- ❑ Use a classification scheme to describe the contents of each ballot rather than having reviewers decide whether or not it is a vote. The reviewers conducted an abstracting process in which they described whatever was marked in the presidential and senate areas of the ballot with a coding scheme. The coding scheme was designed in a systematic process. Each of the prevalent voting technologies in Florida was studied to see how it worked, and how it could fail. Codes were then designed to describe each of those potential failures. Field tests were conducted with each of the major technologies, Votomatic punchcards, Datavote punchcards and paper ballots fed through optical scanners. Field testers were interviewed about any difficulty in applying the codes or markings not covered by the codes, and the coding schemes were then finalized. Ballot viewers were also instructed to make notes of anything written on the ballots or other oddities, such as torn or scuffed ballots. Academic specialists in design methodology were consulted as the project took shape.

Data Collection

The consortium hired National Opinion Research Center (NORC) at the University of Chicago to perform the field research. The project was under the direction of Vice President Kirk Wolter and Project Manager Diana Jergovic. NORC assigned teams of three ballot viewers managed by a team leader who was an experienced NORC field research employee, most of whom were brought in from around the country. Team leaders were responsible for making sure each reviewer worked independently, as well as for collecting the coding sheets, sending them to Chicago for entry into the database and other administrative roles. Having experienced field researchers on site with each team provided consistency and reliable problem solving. Field managers also oversaw the screening and training of ballot reviewers, which included a questionnaire on political bias used to exclude activists, an eye examination and a standardized training program designed by NORC that included practice ballot review under realistic conditions. NORC was responsible for assembling and documenting the database as a public release file.

The consortium also appointed a Data Analysis Working Group (DAWG) under Fessenden and Keating to make sure each news organization would be prepared to handle the data, and also to gather, organize, document and distribute complementary data. DAWG members were Sharon Crenson of Associated Press, Elliot Jaspin of Cox Newspapers (Palm Beach Post), Connie Humberg of the St. Petersburg Times, Sean Holton of the Orlando Sentinel, Richard O'Reilly of the Los Angeles Times, Archie Tse of the New York Times, Ed Foldessy of the Wall Street Journal and Keating Holland of CNN. The group was assisted by Jergovic, Jane Caplan of CNN and Bob Drogin of the Los Angeles Times. The consortium data included two 67-county surveys. The Associated Press and CNN queried counties about practices on election day, such as whether precinct-level ballot checking was equipment turned on and what ballot design was used. The Florida newspapers asked the county election Canvassing Boards about practices and policies during the statewide recount of undervotes ordered by the Florida

Supreme Court, such as whether they were counting just undervotes or overvotes as well and what marks each county was counting as valid votes.

Holton also assembled precinct-level election results (noting several errors in the state's certified final results). Finally, the group gathered precinct-level demographics for gender, race, party and new registrants.

NORC's data gathering process began at the start of February and lasted through April. Under Florida law, ballot reviewers were not able to touch the ballots. County election officials held the ballots in front of the reviewers for their examination. Based on results from the field studies, the consortium issued a photographer's light box to each reviewing team because consistency of lighting is important in detecting incomplete marks on punchcards. Counts of ballots reviewed were compared to election-day figures for undervotes and overvoted ballots. After review of the count of ballots examined indicated shortcoming in some precincts, NORC teams returned to several counties to re-examine precincts. That process concluded in late May.

NORC conducted an additional re-coding study concurrent with the main project. After undervote ballots for a precinct had been examined, team leaders in the field asked county elections officials to put that precinct aside instead of putting it away. Later in the day, the undervotes for that precinct would be brought back to the ballot reviewers to examine how consistently each reviewer would re-code the same ballots. Reviewers did not know which ballots would be re-tested.

The ballot dataset and ancillary data was a joint effort of NORC and the consortium. Through the DAWG, the consortium also agreed on a set of uniform standards to be applied to the ballots, and scenarios of which ballots would have been counted under various recount possibilities. Those rules are reflected in the media consortium's readme file released with the public data.

All analysis provided for this paper, however, was done by Dan Keating for The Washington Post and does not represent the findings or opinions or work of others unless specifically noted. The author is particularly indebted to Diana Jergovic for her constructive suggestions on the draft of this work.

Results

Some of the revelations were particular to the November 2000 election – who might have won statewide recounts? But there is much fresh evidence for election reform and for improving administration of future elections and recounts: Punchcards show markedly higher error rates. Ballot design has been underrated as a factor in voter error. African-Americans and whites vary in quantity and type of error. Dimples and other errant marks fit the same pattern as valid votes. Recounts involve substantial human subjectivity.

Candidate Outcomes

When uncounted ballots were reviewed for potential votes, two critical findings emerged: The recount outcome did *not* hinge on whether dimples or other incomplete marks were counted as votes. And, because of misjudgments about what was likely on the ballots, both candidates pursued strategies that were diametrically opposite to their best interests during the recount. Any discussion of recount outcomes must note that the media consortium ballot analysis used impartial, multiple reviews of ballots and computerized application of standards, none of which would have happened in an actual hand recount. For that reason, the ballot review is a best approximation of what was on the ballots, but not a firm prediction of what would have happened in a recount.

Rather than dimples or not-dimples, the deciding factor in the recount was inclusion of all ballots or only a subset of ballots. And the deciding line was very simple – if all of the ballots were counted there were enough potential Al Gore votes to give him a victory, but any smaller subset of ballots would retain or even enlarge George W. Bush's margin.

Table 1
Candidate Outcomes Based on Potential Recounts in Florida Presidential Election 2000

Review of All Ballots Statewide (Never Undertaken)		
Review Method	**Winner**	**Margin of Victory**
Standard as set by each county Canvassing Board during their survey	**Gore**	171 votes
Fully punched chads and limited marks on optical ballots	**Gore**	115 votes
Any dimples or optical mark	**Gore**	107 votes
One corner of chad detached or optical mark	**Gore**	60 votes
Review of Limited Sets of Ballots (Initiated But Never Completed)		
Review Method	**Winner**	**Margin of Victory**
Gore request for recounts of all ballots in Broward, Miami-Dade, Palm Beach and Volusia counties	**Bush**	225 votes
Florida Supreme Court of all undervotes statewide	**Bush**	430 votes
Florida Supreme Court as being implemented by the counties, some of whom refused and some counted overvotes as well as undervotes	**Bush**	493 votes
Certified Result (Official Final Count)		
Recounts included from Volusia and Broward only	**Bush**	537

Table 1 shows the candidate outcomes broken out by full or partial inclusion of ballots, and acceptance of different marks. For ballots described in the data by three ballot reviewers, Table 1 uses agreement of two-out-of-three in meeting the standard being applied, as does other analysis in this paper unless otherwise noted.[1]

Since the media consortium's findings indicated that Gore could have won a full statewide recount of all votes (all undervotes and overvotes), there has been debate about the likelihood of such a recount. Judge Terry Lewis, assigned by the Florida Supreme Court to oversee a statewide recount of undervotes – which was terminated by the U.S. Supreme Court – wrote in a note during the recount[2] and has said in interviews[3] that feedback from counties may have led him to order all votes counted. Based on that assertion, some Democratic advocates have said that Gore would likely have won the statewide recount that was underway. Needless to say, in a dispute that was litigated at every step – twice to the U.S. Supreme Court – any changes considered by Judge Lewis would not have passed without scrutiny.

Ironically, however, the Republicans argued for changes that could have undercut a Bush victory. It was the Bush's attorneys who argued before the U.S. Supreme Court that leaving overvotes out of the statewide recount did not provide equal protection, allowing voters who erred by undervoting a second chance denied to voters who overvoted. What the Republicans did not know was that the overvotes could yield the cache of votes needed for Gore to overturn the election. That misunderstanding followed

[1] The media consortium Data Analysis Working Group document describes in length "general agreement" meaning that reviewers agree that a standard was met, as compared to "precise agreement" in which they agree exactly on what mark is present on each chad. It also describes unanimous agreement among all reviewers and two-out-of-three agreement. Given two forms of "agreement" and two levels of agreement, there are four potential outcomes for reconciling multiple views of each ballot. Because precise agreement is irrelevant as to whether a ballot meets a given standard of granting votes (i.e. at least a dimple or at least one-corner detached), general agreement is used. Because unanimous agreement means sometimes agreeing with one viewer in opposition to what the other two viewers saw, two-viewer agreement is used.

[2] Newsweek web exclusive article 19 November 2001 cites Lewis' hand-written noted on correspondence during recount indicating that recount of all underovtes and overvotes was necessary.

[3] Orlando Sentinel article 12 November 2001 quotes Lewis in interview saying that recount of undervotes and overvotes was under consideration until the U.S. Supreme Court stopped the statewide recount.

the broader pattern of each party employing strategy that contradicted its best interest during the recount controversy.

The findings from the ballots are in stark contract to the Gore strategy of pursuing punchcard votes in Palm Beach, Broward and Miami-Dade and arguing for inclusive standards on undervoted cards. Conversely, Bush would have gained from encouraging punchcard recounts, which Republicans contested in the South Florida counties. And the assertion by his lawyers that overvotes had to be included in a recount would have been disastrous for him. Because no one had ever looked at ballots the way the Florida Ballot Project did, people guessed wrong about how the pattern of failed votes would fall out.

Table 2
Yield of Presidential Recount Votes by Technology and Undervote-Overvote,
Using Dimple Punchcard Standard and All Candidate Marks on Optical Ballots,
Florida 2000

Undervotes		
	Technology	**Net Gain**
	Votomatic Punchcard	Bush 416 votes
	Datavote Punchcard	Bush 24 votes
	Optical Ballots	Gore 140 votes
	Total	**Bush 300 votes**
Overvotes		
	Technology	**Net Gain**
	Votomatic Punchcard	Gore 225 votes
	Datavote Punchcard	Bush 2 votes
	Optical Ballots	Gore 662 votes
	Total	**Gore 885 votes**
Undervotes and Overvotes		
	Technology	**Net Gain**
	Votomatic Punchcard	Bush 191 votes
	Datavote Punchcard	Bush 26 votes
	Optical Ballots	Gore 802 votes
Total		**Gore 585 votes**

Technology Impacts on Voter Error

The review of uncounted ballots revealed patterns in how many failed ballots could be retrieved, and how that varied between voting technologies. The proportion of undervotes was uniform across different punchcard and optical ballot technologies – about 1-in-3 undervoted ballots could potentially become a vote in a recount. Overvotes showed a much greater variation. Less than two percent of punchcard overvotes could be converted into votes in a recount. But overvoted optical ballots often had errors that left potential votes to be claimed in a recount. The most common was the "double-bubble" first revealed by the Orlando Sentinel's recount in Lake County.[4] On those ballots, a voter marks a candidate and then where it says "Write In Candidate," the voter follows the instructions and writes the candidate's name. If both the candidate and "write-in" ovals are filled (or arrows completed), machines read an overvote and presidential choice

Table 3
Ballots That Could Be Reclaimed in a Recount By Technology and Undervote-Overvote, Dimple Standard on Punchcards or Any Candidate Mark on Optical Ballots, Florida 2000

	Technology	Ballots Reviewed	Potential Votes	
Undervote				
	Votomatic Punchcard	53,215	18,351	35%
	Datavote Punchcard	771	259	34%
	Optical Ballots	6,865	2,314	34%
	All	**60,851**	**20,924**	**34%**
Overvote				
	Votomatic Punchcard	84,822	641	1%
	Datavote Punchcard	4,427	80	2%
	Optical Ballots	24,400	3,008	12%
	All	**113,649**	**3,729**	**3%**
All		**174,500**	**24,653**	**14%**

[4] Story 19 December 2000 shows votes could have been reclaimed in hand review of overvoted ballots.

is invalidated. Since the write-in was not a "qualified write-in candidate," state election law[5] says that the write-in is void and the vote counts for the candidate chosen. Several counties review those votes on election night as standard procedure.

Examining the uncounted ballots divided those contested ballots into more important categories: ballots on which the voter made no mark whatsoever – apparently genuine undervotes, ballots on which the voter made some apparent attempted vote that might have been retrieved in a recount and ballots on which the voter marked more than one candidate, invalidating the ballot. It would be reasonable to hypothesize that voting technology would have no influence on the likelihood of voters having no opinion in the presidential race. The findings contradicted that supposition, showing that punchcards

Table 4
Ballot Error Rate per 1,000 Ballots Cast by Technology
Differentiating Between Genuine "No Opinion" Undervotes and Failed Votes, Florida 2000[6]

Technology	Counties	Ballots Cast	Failed Mark For One Candidate	No Attempted Mark "Genuine Undervote"	Marks for More Than One Candidate	**Total**
Votomatic Punchcard	15	3,642,825	6.7	7.8	23.4	**37.9**
Datavote Punchcard	9	76,609	2.8	5.9	56.1	**64.8**
Optical Ballots	16	401,546	6.7	3.2	40.2	**50.0**
Optical Ballots with Precinct Checking	25	1,951,700	1.4	1.7	2.7	**5.9**

[5] Florida Administrative Code: 1S-2.0031 **Write-in Procedures Governing Electronic Voting Systems** (7) An overvote shall occur when an elector casts a vote on the ballot card and also casts a write-in vote for a qualified write-in candidate for that same office. Upon such an overvote, the entire vote for that office shall be void and shall not be counted. However, an overvote shall not occur when the elector casts a vote on the ballot card but then enters a sham or unqualified name in the write-in space for that same office. In such case, only the write-in vote is void.

[6] The differences between each of the technologies for each type of error and overall error is statistically significant beyond the 99.9% confidence level (p < .001).

had much higher rates of genuine undervotes – no opinions expressed at all. Voters using Votomatic punchcards were four times as likely to express no opinion whatsoever than voters using optical paper ballots. That may indicate that punchcard technology is hard for some people to use. It also may be affected by the ease of making some kind of mark – including a protest mark – when the voter is holding a pencil and using a paper ballot. Table 4 shows that Datavote punchcards – which use a spring-launched hole puncher – are much less likely than Votomatic punchcards to have partial or incomplete marks. But their overvote rate is twice as high, possibly because of challenges in lining up the bulky sliding punch mechanism on Datavote machines. So the technology trades away the dimple problem but produces a greater overvote, yielding a higher error rate overall.

Ballot Design and Instruction Issues

The first technical flaw to draw attention in the Florida election was the confusingly designed Palm Beach "butterfly" ballot, on which candidate names were interspersed on two sides of the page with the punchholes down the center. Attention to incomplete marks on punchcards (dimples and hanging chads) then turned attention to problems with punchcards as a whole. Criticism of dimples and the butterfly ballot may have obscured the broader issue of ballot design failures. The challenge of fitting 10 president-vice president pairings onto a ballot provided a rich experiment in how defective designs affect votes. There were defective and effective designs in punchcards and with optical paper ballots.

In contrast to the beleaguered punchcards, precinct-level ballot checking as used in some of Florida's optical ballot counties drastically reduces voter error rates, as shown in Table 4. But the results found that even the precinct checking was no match for defective ballot designs. Counties with second-chance technologies, but saddled with misleading ballots, had much higher error rates than counties with error-prone technology but clear ballots. For optical paper ballots or Datavote punchcards, the defective designs were called "caterpillar" because seven or eight candidates were listed in the leftmost column of the ballot, and the remaining candidates were listed in the next column to the right. Many voters marked one candidate in each column. In Votomatic counties, Palm

Table 5
Ballot Design Flaws Exceed Technology as Cause of Voter Error,
Rate of Errors Per 1,000 Ballots Cast, Florida 2000[7]

	Ballot Design	Counties	Ballots Cast	Failed Mark For One Candidate	No Attempted Mark "Genuine Undervote"	Marks for More Than One Candidate	Total
Votomatic Punchcard	No Design Prob	13	2,888,056	4.8	8.2	15.3	**28.3**
	Design Prob	2	754,769	13.9	6.5	54.3	**74.6**
Datavote Punchcard	No Design Prob	7	42,900	2.6	6.3	57.3	**66.2**
	Design Prob	2	33,709	3.1	5.3	54.6	**63.1**
Optical Ballots	No Design Prob	2	212,993	6.7	3.4	26.0	**36.2**
	Design Prob	14	188,553	6.6	3.0	56.2	**65.7**
Optical Ballots with Precinct Checking	No Design Prob	25	1,951,700	1.4	1.7	2.7	**5.9**
All	No Design Prob	47	5,095,649	3.6	5.5	11.3	**20.4**
	Design Prob	18	977,031	12.1	5.7	54.6	**72.5**

[7] Overall and within each technology, the difference between places with and without ballot design flaws is significant beyond the 99.9% confidence level (p < .001).

Beach's butterfly was joined by Duval County's punchcard that listed five candidates on the first page and five candidates (and the write-in spot) on the second page when voters are instructed to "vote each page."

Overall, the error rate for counties with bad ballot designs is more than 3.5 times as high as for counties without design problems. Votomatic showed the greatest increase, followed by Optical technology with centrally tabulated ballots. Datavote's error rate was actually slightly lower with the defective design. No counties with optical precinct checking had defective ballot designs.

Racial Disparities

Across all technologies, African-American voters had higher rates of ballot errors than whites. Second-chance technology, however, reduced the discrepancy in error rates between whites and African Americans, from a factor of 2.5 times as high to 1.8 times as high, indicating that African Americans may be the biggest beneficiaries of election reform that have required precinct-checking technology. Ballots do not give any indication of the voter's age, race, sex or other identifying information. Racial analysis was done by using precincts that were either 80% white or 80% African-American.

African Americans and whites had different patterns of errors. The two groups had a similar likelihood of erring with an attempted vote for one candidate that resulted in an undervote, such as by making a dimple. African Americans, however, were more than twice as likely to make no mark whatsoever. African Americans were more than five times as likely to overvote, with the difference showing up most notably on

Table 6
Voter Error Rates per 1,000 Ballots Cast By Race and Technology,
Race Assigned by Precincts 80 Percent African-American or White, Florida 2000[8][9]

Technology	Race	Precincts	Ballots Cast	Attempted/ Failed	No Attempt - Genuine Undervote	Over vote	All
Optical Ballots	**African American**	7	3,186	12.6	6	140.3	**158.8**
	White	271	197,222	7.3	2.2	37.7	**47.2**
Optical Ballots w/ Precinct Check	**African American**	1	27,622	3.2	2.9	7	**13.1**
	White	1,107	1,153,382	1.4	1.8	2.2	**5.4**
Votomatic	**African American**	187	152,453	7.4	13.8	89.8	**111.0**
	White	2,021	2,114,362	6.1	6.1	17.5	**29.6**
All	**African American**	235	183,261	6.9	12.0	78.2	**97.1**
	White	3,399	3,464,966	4.6	4.4	13.6	**22.6**

Votomatic punchcards. Those patterns coupled with tendencies of whites to vote for Bush and African Americans to vote for Gore[10] help explain why a recount of punchcards to retrieve undervotes could have helped Bush more than anticipated, while including overvotes in a recount would have helped Gore.

Just as Table 6 showed that technology influenced differences in error rates between the races, ballot design issued played a role, as well. Bad designs worsened error rates under each technology with either race. African Americans error rates more than doubled with both Optical and Votomatic technologies with bad designs. The greatest impact on whites was in Votomatic punchcard counties, where the error rate more than doubled.

[8] Counties using Datavote punchcards, Hand and Lever technology excluded because there were too few precincts identified as 80 percent African American for analysis.

[9] The differences expressed in error rates between races across technologies are all significant beyond the 99.9% confidence level (p < .001).

[10] Exit polls showed Gore getting more than 90 percent of African-American votes while Bush carried a majority of white votes.

Table 7
Error Rates per 1,000 Ballots Cast by Technology and Ballot Design and Race,Race Assigned by Precinct 80 Percent African-American or White, Florida 2000

Technology	Race	Ballot Design	Precincts	Ballots Cast	Attempted/ Failed	No Attempt - Genuine Undervote	Over vote	All
Optical Ballots	**African-American**	**No Design Prob**	1	201	24.9	0	49.8	**74.6**
		Design Prob	6	2,985	11.7	6.4	146.4	**164.5**
	White	**No Design Prob**	75	72,407	9.4	1.5	24.3	**35.2**
		Design Prob	196	124,815	6.1	2.6	45.5	**54.2**
Optical Ballots w/ Precinct Check	**African-American**	**No Design Prob**	41	27,622	3.2	2.9	7	**13.1**
	White	**No Design Prob**	1,107	1,153,382	1.4	1.8	2.2	**5.4**
Votomatic punchcards	**African-American**	**No Design Prob**	128	109,010	6.1	15.0	61.2	**82.3**
		Design Prob	59	43,443	10.7	10.7	161.7	**183.0**
	White	**No Design Prob**	1,511	1,597,674	3.9	6.3	10.2	**20.4**
		Design Prob	510	516,688	12.9	5.3	40.1	**58.3**
All	**African-American**	**No Design Prob**	170	136,833	5.6	12.5	50.2	**68.3**
		Design Prob	65	46,428	10.7	10.4	160.7	**181.8**
	White	**No Design Prob**	2,693	2,823,463	3.0	4.4	7.3	**14.6**
		Design Prob	706	641,503	11.6	4.7	41.2	**57.5**

Badly designed ballots had the effect of narrowing the gap in error rates by race slightly, as error rates for whites went up by a factor of almost three with bad ballot designs, while error rates for African-Americans increased by a factor of less than two.

Overvoting

Overvoting – marking more than one candidate – was, by far, the most common form of failed voting with all technologies. As shown in Table 5, overvoting was drastically reduced by second-chance precinct counting technology and non-defective ballot designs.   As Tables 6 and 7 showed, African Americans were more likely than whites to overvote.

The overvoting also showed a very notable difference in which candidates were named on the overvoted ballots. Gore's name appeared on 80,772 of the overvotes compared to 40,073 for Bush, indicating that overvotes may have had the largest impact on Florida's election. Voters included both Bush and Gore on 11,409 overvoted ballots. Only 4,384, or 3.9 percent, of overvotes had neither Bush nor Gore included.

Defective ballot designs alone may have cost Gore many more votes than the final 537-vote margin of the election. Palm Beach County's butterfly ballot design yielded 8,170 voters who overvoted by punching Gore and one of the candidates who flanked him, Patrick Buchanan or David McReynolds. Another 1,668 voters punched Bush and Buchanan, the only name flanking Bush. The net effect of those errors cost Gore 6,502 votes.

A similar effect took place in Duval County. The heavily Republican county had a ballot spoilage rate twice as high as Palm Beach, which was attributed to spreading the presidential candidates' names over two pages when voters are instructed to vote on every page. Examination of voters who chose just Bush or just Gore on the front page and one additional candidate from the second page indicates that Bush lost 4,465 votes from that error and Gore lost 7,050, a net loss for Gore of 2,585.

Neither the Palm Beach butterfly overvotes nor the Duval two-candidate two-page overvotes were awarded to any candidate in the media consortium's review since there was no indication that votes could ever be awarded in that situation in a recount.

Validity of Incomplete Marks

An issue that arises in every recount is what kind of failed votes can be counted. Some advocate a standard that measures by how closely voters followed the instructions ("push through the card") while others use discernable voter intent as the yardstick. Florida's debate over "Do dimples count?" was passionate and partisan, but the conflict is not inherently partisan. Democrats and Republicans have argued on both sides of the debate in other recounts.

To test the question of whether incomplete marks should be considered votes, the marks were analyzed to see if they fall out in a random pattern or synchronize with the other votes on the ballot. Analysis of this question was done with the media consortium Florida Ballot Project data and also with a collection of more than 3 million ballots from the ballot image files created by ETNet tabulation software in some Votomatic punchcard counties. The ballot image files have one record for each ballot and one field for each candidate position (chad), showing whether the punchcard reader detected a punch or no punch. The ballot image files allow analysis of behaviors for a given voter across the ballot, including patterns of punches on overvotes.

To establish a baseline, 3 million valid votes were examined and categorized to see what percent of voters chose the same party in the presidential and U.S. Senate races (the only two races that were consistent statewide). It showed that 86 percent of Democratic presidential voters and 83 percent of Republican presidential voters chose the same party in the Senate race.

Failed undervotes that had only one marking in the presidential area were then examined for consistency between the party of the presidential mark and the party of the Senate vote. The failed votes were broken out by technology and types of failure, such as dimple-or-1-corner detached, two- or three-corners detached, all corners detached[11] , wrong ink-color on optical ballots, incomplete marks on optical ballots and marks

Table 8
Pattern of Party Matching Between Presidential and U.S. Senate Votes
Comparing Valid Votes to Incomplete Marks by Technology and Voter Error and Party, Florida 2000

Technology	Presidential Markings	Presidential Party	Same party in Senate	Not Same Party in Senate	Percent Same Party President-Senate
Punchcard Ballots	Valid Votes	Democrat	1,377,455	230,409	**86%**
		Republican	1,070,438	223,701	**83%**
	Fully Punched Chad Undervote	Democrat	205	61	**77%**
		Republican	280	100	**74%**
	Two-corner or Three-Corner Detached Chad	Democrat	220	88	**71%**
		Republican	481	162	**75%**
	Dimple or One-Corner-Detached Chad	Democrat	6,181	1,386	**82%**
		Republican	5,996	1,588	**79%**
Optical Ballots	Wrong Ink Color	Democrat	265	57	**82%**
		Republican	366	108	**77%**
	Underfilled Oval	Democrat	177	60	**75%**
		Republican	106	50	**68%**
	Mark Away from Oval	Democrat	474	143	**77%**
		Republican	284	108	**72%**

[11] Fully-punched chad undervotes are ballots on which a chad was apparently pressed back into an empty hole as the cards went through the punchcard readers, or in which the chad was missing entirely but the ballot had been characterized as an undervote.

elsewhere, such as circling the candidate name. Table 8 shows the pattern of party consistency in valid votes and incomplete marks, with both demonstrating a strong correlation between the presidential party and senate party. The correlation beyond the 99.99% confidence level (p < .0001) signals that the dimples are not random acts.

The incomplete presidential marks fall out in the same pattern as normal presidential votes in terms of party consistency. Failed votes mimicked the party-consistency pattern of normal presidential votes, indicating that dimples aren't random acts, but are most likely attempted votes.

Subjectivity of Manual Recounts

On the morning after the Nov. 7 election, Tampa election supervisor Pam Iorio made clear why rumors of a statewide hand recount frightened her: Letting the machines count applies a consistent – even if imperfect – standard everywhere, while a hand recount injects millions of subjective judgments into the process.

By having three people independently judge each undervote ballot under close supervision, the Florida Ballot Project measured the subjectivity of looking for marks on ballots. Does everyone see the same thing? Do different people see different things?

The Florida Ballot Project data includes a ballot reviewer identification (or "coder ID") for every set of descriptive codes (up to three per ballot, all in one record). The identifications link to a separate table of demographics for the ballot reviewers that was collected voluntarily after they were hired, including age, gender, party affiliation, party of vote for president in 2000, income and education level.

NORC performed preliminary analysis on coder reliability for the media consortium as part of its contract. The first analysis done was of consistency on overvoted ballots based on the three counties where three coders reviewed all overvotes as well as undervotes. NORC compared the three coders on a chad-by-chad basis and found that in Polk County, which used optical technology, the reviewers examined 117 ballots and had a level of agreement of .975, with 1.0 signifying total agreement. The analysis was done pairwise with 117 ballots with 10 presidential positions and three reviewers. So each reviewer has 1,170 codes that yield 2,340 comparisons with the other two coders. There were a total of 86 pairwise disagreements, yielding the .975 agreement rate. In Nassau County, a Datavote punchcard county, NORC examined 331 ballots and the overall level of agreement was .985. In Pasco County, a Votomatic punchcard county, NORC had 744 ballots. The level of agreement was .99. Examining all the candidate choices as a group instead of position-by-position, NORC found all three reviewers agreed 92 % in Polk County, 95% in Nassau County and 92% in Pasco County. Based on those findings, the consortium switched to using only one ballot reviewer for overvotes, keeping three reviewers for undervotes because of the finer issues of discernment in describing incomplete marks as compared to multiple marks that were clear enough to be read by a machine.

NORC preformed preliminary analysis of coder reliability based on various demographic factors. NORC released some findings to the media consortium and is preparing a paper for publication. NORC performed pairwise comparison on a position-by-position (chad-by-chad) basis and found 99.1% agreement on optical paper ballots, 96.5% agreement on Datavote punchcards and 95.9% agreement on Votomatic punchcards.

NORC analyzed different potential vote standards for differences in reviewer agreement. On punchcards, NORC re-coded each chad position based on whether the reviewer saw a least a dimple ("dimple or greater mark") or whether the coder saw at least two corners of a chad detached ("two-corner or greater mark"). On optical aper ballots, NORC recoded each presidential position as to whether the reviewer had seen any candidate mark including circling the name or the party ("any mark for candidate") or whether the reviewer saw some mark specifically on the oval or arrow ("mark by oval/arrow for candidate"). Table 10 shows that reviewers agree more often on optical ballots and agree more often on the "Other" candidates than on the Bush or Gore positions, most likely because the Other candidates are almost always blank. Using the more restrictive standards – two-corner detached or a mark around the oval/arrow – also increased consistency between reviewers.

Table 9
Rough Summary of NORC Analysis of Reviewer Consistency, Pairwise, Florida 2000

	Pair comparisons	Disagreements	Agreement Rate
Votomatic Punchcards	319,041	13,081	0.959
Datavote Punchcards	4,300	150	0.965
Optical Ballots	43,000	400	0.991

Table 10
Summary of NORC Analysis of Reviewer Consistency For Different Vote Standards on Different Technology By Presidential Position, Florida 2000

Technology	Standard Applied	Bush Chad/ Position	Gore Chad/ Position	Other Candidates
Votomatic	dimple or greater mark	88%	87%	99%
	two-corner or greater mark	98%	99%	100%
Datavote	dimple or greater mark	79%	86%	99%
	two-corner or greater mark	87%	92%	100%
Optical	any mark for candidate	96%	94%	99%
	mark by oval/arrow for candidate	98%	98%	100%

In searching for factors correlating to reviewer disagreements, NORC found no significant differences based on demographics for optical or Datavote ballots. But it found significant differences for both gender and party on Votomatic ballots, with men more likely to find marks on either the Bush or Gore chads than women.

Table 11
Summary of NORC Analysis of Gender and Party Influence on Coder Findings on Votomatic Ballots, Florida 2000

	Overall Odds	Male Reviewer Odds	Female Reviewer Odds	Republican Viewer Odds	Democrat Reviewer Odds
Bush chad	0.180	0.216	0.144	0.202	0.158
Gore cad	0.213	0.220	0.206	0.209	0.217

Republicans were more likely than Democrats to find marks on the Bush chad, and vice versa. The impact of both was greater on the Bush chad.

NORC also released to the media consortium preliminary information on the ballot reviewer re-coding study, in which reviewers were shown the same set of ballots at two different times to see how often a ballot reviewer would agree with his or her original work. NORC said the overall self-agreement rate was 98.3%, with a range for each reviewer from 93.2% to 100%. No details were available for technology.

The Post conducted separate analysis of reviewer consistency intended to be more directly applicable to impacts on recounts. Studies of agreement at the chad- or position-that drives consistency in a recount, not necessarily agreement on the lack of marks. For level have inflated agreement because the vast majority of the 10 presidential positions are blank the vast majority of the time, as was shown by NORC finding higher agreement on the "Other" positions, which are least frequently marked. It is agreement on the marks instance, if two people look at a ballot with one person seeing no marks and the other person seeing a single candidate marked, a chad-by-chad analysis would find 90 percent

Table 12
Disagreement in Potential Votes Seen Between One, Two and Three Ballot Reviewers,
Dimple Standard or Any Candidate Mark on Optical Ballots
Florida 2000

Reviewer Sees Valid Vote Pattern	Bush Vote	Gore Vote	Margin for Bush	Total Votes	Disagree Group	Disagree Pct
Votomatic Punchcards						
One Sees Vote (Y-N-N)	1,612	1,808	-196	3,420	7,840	**36.9%**
Two See Vote (Y-Y-N)	2,057	2,363	-306	4,420		
Three See Vote (Y-Y-Y)	7,068	6,345	723	13,413		
Datavote Punchcards						
One Sees Vote (Y-N-N)	22	16	6	38	83	**27.4%**
Two See Vote (Y-Y-N)	24	21	3	45		
Three See Vote (Y-Y-Y)	122	98	24	220		
Optical Ballots						
One Sees Vote (Y-N-N)	27	59	-32	86	199	**8.1%**
Two See Vote (Y-Y-N)	45	68	-23	113		
Three See Vote (Y-Y-Y)	1,036	1,215	-179	2,251		
All						
One Sees Vote (Y-N-N)	1,661	1,883	-222	3,544	8,122	**33.8%**
Two See Vote (Y-Y-N)	2,126	2,452	-326	4,578		
Three See Vote (Y-Y-Y)	8,226	7,658	568	15,884		

agreement, while a vote-level review would find complete disagreement.

To test consistency on a vote level, the Post tested ballots screened by three reviewers on which at least one of them saw a pattern in the presidential area that would constitute a vote for Bush or Gore using the dimple standard. Of those 24,006 ballots, all three reviewers saw a pattern constituting a vote 15,884 times. On the other 8,122 occasions, two people saw one thing and the other person saw something else. That is 33.8% disagreement from the universe of potential votes.

While most people could agree that nothing was marked on most of the undervotes, discerning which of the partially marked ballots had potential votes proved very subjective. Ballot viewers disagreed about one-third of the potential votes for Bush or Gore.

The disagreement rate was much higher in punchcard technologies. Consistency was much greater in viewing the paper optical ballots, which, unlike punchcards, are designed to be read directly by voters and are clear to the naked eye. Using a more restrictive standard for assigning votes, such as two-corners of a chad detached, did not lower the rate of disagreement; the size of the "agreement" group declined more than the "disagreed" group. The share of potential votes with disagreement rose on Votomatic to 43.4% and on Datavote to 52.6%.

Table 13
Disagreement in Potential Votes Seen Between One, Two and Three Ballot Reviewers,
Two-Corner Chad Or Greater Mark,
Florida 2000

Reviewer Sees	Bush	Gore	Margin for	Total	Disagree	Disagree
Valid Vote Pattern	Vote	Vote	Bush	Votes	Group	Pct
Votomatic Punchcards						
One Sees Vote (Y-N-N)	299	242	57	541	875	**43.4%**
Two See Vote (Y-Y-N)	192	142	50	334		
Three See Vote (Y-Y-Y)	759	384	375	1,143		
Datavote Punchcards						
One Sees Vote (Y-N-N)	25	20	5	45	91	**52.6%**
Two See Vote (Y-Y-N)	24	22	2	46		
Three See Vote (Y-Y-Y)	45	37	8	82		

The study methodology and analysis were not designed to mimic a Canvassing Board, in which three people share their opinions of what they see on a ballot and reach a consensus or decide by a 2-1 margin. The demographic influences found by NORC on Votomatic ballots and subjectivity detected in the Washington Post analysis could be smoothed out as three people confer about a ballot. The study shows, however, that discerning what is on a ballot is a subjective process influenced by various factors, especially on punchcards.

Discussion

There are almost endless questions that can be asked of the ballots at issue in the November 2000 Florida election for president. Many times during the 36-day recount process, scholars and participants harkened to the 1876 presidential contest that was thrown into the House of Representatives because of a disputed electoral delegation from Florida. In similar fashion, what happened and what was learned from Florida 2000 may be a guidepost for centuries to come. The media consortium's Florida Ballot Project data will allow anyone to delve into questions of how it came about and how it could have been different. The supplemental data, including the ballot image files, help to provide a rich source of investigation.

The need to heed what the data tells us is shown by how both presidential campaigns struggled to find effective recount strategies – and how both ended up on tacks that were not to their own advantage. Al Gore ignored the rich cache of optical ballot overvotes, a set of ballots in which clearly legal votes were counted in some counties and ignored in others. Meanwhile, George W. Bush's campaign resisted recounts of punchcard undervotes that would have helped cement his victory.

For political activists, the enduring lesson of Florida is that voters not only have to be shepherded to the polling booth, but have to trained what to do when they get there. The notorious Voter News Service poll that led broadcasters to predict a Gore victory might well have accurately gauged the intention of voters in the booth, but was misled by the skewed nature of overvoting that apparently cost Gore tens of thousands more votes than Bush. Many of the overvotes can be attributed to confusion caused by the write-in position on the ballot. By contrast, only 40 write-in votes were included in the final certified results. Obviously, the write-in system cost thousands of voters their voice for virtually no benefit in voter choice.

Elections reformers need to avoid knee-jerk rejection of punchcard ballots. Though precinct-level checking was only employed with optical ballot systems in Florida, technology is available to check punchcards in the polling place, as well. Checking in the polling place had a dramatic effect on reducing ballot error and the discrepancy between African-American and white error rates. African Americans were more than twice as likely as whites to make no mark whatsoever in the presidential area of their ballot. One possible explanation could be that Gore won more than 90 percent of African-American votes, indicating that Bush had almost no support. Voters in that community who found Gore unappealing might have preferred to leave a blank ballot than to vote for the major-party opponent. The higher rate of undervotes with no mark whatsoever on punchcards than optical ballots may indicate that some technology is so bewildering that people are unable to make any attempt at registering their vote.

One of the most surprising and critically important findings was that bad ballot designs can be as important as error-prone technologies in causing voter error. Confusing caterpillar, butterfly and two-page ballots led to much higher error rates. Spending hundreds of millions of dollars on new technology will not prevent voter errors if ballots (or computer screens) are confusing, or if voters are not given accurate guidance. Attention must be paid to ballot design and testing rather than assuming that spending for new technology will solve the problems. Pre-testing with a small but representative sample of county residents would be an easy, low-cost and effective method of detecting and correcting this kind of ballot design fiasco.. Voter education, clear instruction and available help at the polling place would also reduce those voter errors.

Finally, counties must apply uniform rules for checking uncounted ballots such as double-bubble overvotes, which were converted to votes on election night in some places and remained uncounted elsewhere – an inconsistency in execution and lack of fairness that will not disappear with new equipment.

The findings may resolve the long debate over whether dimples should or should not count during a recount. All incomplete marks followed a pattern of party-consistency that resembled valid votes, refuting the assertion that dimples are random acts and bolstering the idea that they are attempted votes that reflect the intent of the person who caused them.

That validity of recounts overall, however, may deserve further reconsideration. Having three people independently record what they saw on each ballot showed that subjective judgments play a significant role in recounts. The results contradict claims that incomplete marks are easy to discern. Interpersonal dynamics – subject to political and gender bias – then come to play. While mechanical ballot readers may miss some valid votes, they are intended to be ruthlessly consistent in applying the standards they use.

Appendix

Analysis based on the media consortium Florida Ballot Project data produced under contract by National Opinion Research Center (NORC). Project information available online at www.norc.org/fl with the data specifically at www.norc.org/fl/results. The main dataset is in two forms, raw and aligned. The raw set includes descriptive codes for each chad for the entire presidential and senate areas of punchcard ballots, recorded by position (chad) number and including positions not assigned to any candidate. The aligned version uses only the positions (chads) assigned to candidates and identifies them by candidate name rather than position number. It also contains a number for the count of marks recorded on the unassigned positions. The datasets have one record for each ballot reviewed, 175,010 ballots. It also has a codebook and data dictionary. Supplemental tables include coder demographics, extra ballots reviewed in Orange County, the recode ballot study data and a dataset of written comments by ballot reviewers about any text or special characteristics of the ballots.

Supplemental data produced by the media consortium's Data Analysis Working Group includes ballot-level, precinct-level and county-level tables . The ballot-level table is the consortium's consensus decisions on applying the written comments to the original ballot positions. The precinct-level data has statewide certified results and voter demographics. The county-level data has information from two statewide surveys on policies and practices on election day and during the aborted statewide recount. The ballot-level data is linked to the NORC ballot data by the unique "balnum" identifier. The precinct- and county-level data can be linked to the NORC ballot data by county fips code and precint name.

Further analysis was conducted with the ballot image file with records of 3 million votes cast on Votomatic punchcards and tabulated with ETNet software, which retains a record for each ballot indicating which positions were detected as punched. The dataset covers 10 counties but was not preserved perfectly in all counties. The dataset is currently available from Dan Keating at The Washington Post, but may be stored in a public archive.

All analysis for tallying votes excludes Volusia County because a recount was complete and certified in that county. There was no way to distinguish which ballots had or had not been counted as votes from the set of ballots presented for media review, so there was no way to know if ballots would be double-counted. Because the recount was certified in Broward County, the Florida Ballot Project used ballots that were judged to be undervotes or overvotes during the recount, avoiding possibly double-counting of votes granted during the recount. On the other hand, the Ballot Project used a machine segregation of ballots in Palm Beach County because the recount in that county was never certified.

All analysis for tallying votes also excludes data from reviewer 75683, for whom NORC found an unacceptable level of inconsistency with other reviewers. The reviewer looked at only 80 ballots before being terminated in the field. The reviewer's data is included in the dataset, but was omitted from all analysis based on NORC's recommendation. That was the only reviewer for whom NORC found an indication of a bias toward any candidate, though Washington Post review of the ballot reviews shows inconsistency that would grant votes to both Bush and Gore in a way that indicates the coder simply was not up to the job.

Analysis of technologies used the baltype2 variable in the NORC ballot dataset, which categorizes ballots by the technology used in that county: Votomatic, Datavote or Optical. Analysis of ballots retrieved in a recount by undervote or overvote uses NORC's baltype1 variable, which indicates undervote or overvote. Analysis differentiating which optical counties used precinct-checking use the "central or precinct tab" variable recoding Escambia to 2 because precinct-checking of ballots for errors was turned off in that county. Analysis involving defective ballot designs uses the two_columns variable in the DAWG county-level table, except that Palm Beach County (fips 99) is recoded to "1" to recognize its defective buttefly ballot design. Analysis differentiating votes as attempted/failed, no attempted mark ("Genuine Overvote") or multiple-candidates marked do not use the baltype2 undervote/overvote variable. The differentiating is based on the evaluation of marks by the reviewers. Assignment of precincts to the categories of 80% white or 80% African American was done with the whiterv (white registered voters percent) and blackrv (black registered voters percent) fields in the DAWG precinct-level database. No precinct demographics are available for absentee precincts.

One issue in this study is the difficulty experienced by counties in attempting to segregate undervote and overvote ballots. The novel plan for segregating undervotes was conceived by Miami-Dade Election Supervisor David Leahy to deal with the time constraint on counting his Votomatic punchcard ballots. He commissioned ETNet to write software to use the punchcard ballot readers to segregate undervotes. The idea of segregating undervotes was adopted by the Florida Supreme Court without any factfinding about the viability elsewhere. That capability was not available for optical ballot or Datavote punchcard counties, or to Votomatic punchcard counties that used different card readers. The Florida Ballot Project encouraged counties to segregate ballots by machine and closely monitored rates of segregated ballots from each county to certified rates of undervotes and overvotes. The project commissioned ETNet to write software for segregating overvotes. Users of optical ballot technology said that differences in humidity or light-sensitivity settings could lead to inconsistency in discerning which ballots were undervotes or overvotes in the certified result. The project ended up including a sample within 1% of the expected total of votes, with most of the variation in Votomatic overvotes, the least likely ballots to yield votes in a recount.

References

Brownstein Ronald, “Decision 2000 News Analysis – Divisions Exposed in Election Will Be Obstacles for Bush,” *Los Angeles Times*, The Times Mirror Company, 14 December 2000.

Damaron David, Campbell Ramsey and Roy Rogyer, “Gore Would Have Gained Votes in GOP Stronghold ‘Overvotes’ Counted Elsewhere,” *Orlando Sentinel*, 19 December 2000.

Damaron David and Salamone Debbie, “Court-ordered Recount Was Guessing Game,” *Orlando Sentinel*, 12 November 2001.

Isikoff Michael, “The Final Word? New documents raise questions about news media's findings on the 2000 presidential election,” Newsweek Web Exclusive, Newsweek Inc., 19 November 2001 .

Von Drehle et al., The Political Staff of The Washington Post, 2001, *Deadlock: The Inside Story of America’s Closet Election*, PublicAffairs.

Voter New Service Exit Poll, Election 2000: America’s Choice, *The Wall Street Journal*, 9 November 2000.