

Problems with "Voting Machines and the Underestimate of the Bush Vote"

November 23, 2004*

Leonard Wayne
3059 Sycamore Ave.
La Crescenta, CA 91214-3741
(818)957-4292
lrwayne@earthlink.net

*11/25/2004: I sent the original version of this report to the Caltech/MIT Voting Technology Project team for comments before release. Team member Charles Stewart, apparently the lead author of the Caltech/MIT report, replied in an email 11/23/2004. Stewart seems to mostly agree with my report: "First, my general comment is that I have very little disagreement with what you write, as far as it goes." However, Stewart did raise four issues. Here are the four issues, and my response to each.

- (1) Stewart objects to the tenor of my report: "I say 'as far as it goes' because the tenor of the report suggests that you still do not understand the reason we released it. That is not your fault, obviously, since it's the job of the author to make things clear." I don't completely understand this criticism. I have focused my report on addressing the one conclusion of the Caltech/MIT report that the media have seized upon most intently, namely: "We conclude that there is no evidence, based on exit polls, that electronic voting machines were used to steal the 2004 election for President Bush." An example of this media focus is the Nov. 19, 2004 Associated Press article by Rachel Konrad, "Academia Still Fixated on John Kerry."
- (2) Stewart notes that my use of the word "corrupt" to describe the data in the Caltech/MIT report could be interpreted more than one way, including to suggest malfeasance, which I do not intend. I agree with this point, and I have added a footnote to the report that makes clear what I mean by the use of this word.
- (3) In my report, I took issue with this quote: "Even when they work well, exit polls are too imprecise to lay against the official count, unless every voter is included in the exit poll." Stewart sent a long "clarification" of what he was trying to convey in this quote. Nonetheless, I have left intact my original criticism, since the Caltech/MIT report says what it says, and has not been amended.
- (4) In the original version of my report, I included a copy of an 11/19/2004 email Stewart sent to CNN and me and cc'd to the rest of the Caltech/MIT team. Stewart writes that he prefers not to have the entire email included in the report (I'm not sure why), but that it is OK if I include quotes from the email. I have modified my report accordingly.

The filename for the original report was vtp.pdf. This updated version is vtp2.pdf.

ABSTRACT

The Caltech/MIT Voting Technology Project (VTP) has released a report titled, "Voting Machines and the Underestimate of the Bush Vote, Version 2," dated November 11, 2004. An examination shows strong statistical evidence that the data upon which the report relies are corrupt¹. The data is corrupted in a way that removes patterns from the data, so it is not surprising that the report finds no patterns in the 2004 presidential election exit poll data.

EVIDENCE OF DATA CORRUPTION

The VTP report includes three plots, which compare state exit poll data with actual state election returns for the 2004 presidential election. Had error bars been included on the plots, it would have been apparent that the exit poll data are not real. The exit poll data match the election returns far too closely.

The VTP report examines 51 state exit polls. The standard error on each poll's estimate of the vote for George W. Bush is $\sqrt{p(1-p)/N}$, where p is Bush's true vote, and N is the number of samples in the survey. The data I use for p and N comes from CNN and is given in Appendix I. The calculation shows that for all 51 exit polls, the standard error is never less than 0.93%². On any of the three plots in the VTP report, draw a horizontal line at +0.93% and another at -0.93%. We generally expect about 68% (one sigma) of the 51 exit polls to fall between these two lines. $0.68 * 51 = 35$, so we expect about 35 of the 51 exit poll results to fall between the lines. However, 45 of the exit poll results fall between the lines³. The exit poll data match the election returns much more closely than is allowed by fundamental statistics.

To quantify the severity of this discrepancy, I ran a simple Monte Carlo simulation. I generated 51 random numbers, normally distributed. I recorded how many of these 51 random numbers were less than or equal to one sigma away from the mean. I then repeated the experiment many times (300,000) to see how often 45 or more of the 51 random numbers ended up clustered within one sigma of the mean. The odds that normally distributed data could by chance be so tightly clustered as in the VTP figures is less than 1 in 1,000. Appendix II has the Matlab code used for the simulation.

¹ The word "corrupt" is used here in the same sense it would be for data on a computer disk accidentally exposed to a magnet. No nefarious overtones are implied.

² There are a number of reasons that the percent uncertainty could be *larger* than this. For example, a "cluster sampling error" is sometimes incorporated to account for the fact that exit polls are only conducted at certain precincts, and the demographics at those precincts may not be representative of other precincts. But there is no way the (one sigma) uncertainty can be smaller than 0.93%.

³ By eye it is difficult to determine exactly how many of the data points in VTP Figure 1 (for example) lie within the +/-0.93% boundary. However, I was able to determine the xy coordinates of the data points in the VTP figure by capturing a snapshot of the VTP figure and using the Matlab pointer to determine the pixel coordinates of the data points, and converting these pixel coordinates into the graph's coordinates.

The Matlab simulation is a generous calculation that surely underestimates the severity of the discrepancy⁴, since it assumes the uncertainty for each exit poll is the same as that for the exit poll with the smallest uncertainty, and since it assumes there are no other sources of uncertainty in the exit poll data other than that due to fundamental counting statistics.

The statistical evidence therefore strongly suggests that the data used as a basis for the VTP report is corrupt.

Below I provide a possible explanation for how the VTP data came to be corrupt. But to introduce that explanation, I first include a discussion of how the VTP data cannot be publicly verified.

CALTECH/MIT DATA CANNOT BE PUBLICLY VERIFIED

The VTP authors seem to concede that the data that is the foundation of the report cannot be publicly verified.

Footnote 2 of the report says, in part:

The exit poll data were taken from the cnn.com web site. The poll data can be accessed through <http://www.cnn.com/ELECTION/2004/pages/results/index.html>.

However, the data from this link do not match the data in the VTP report. Details of the discrepancy are contained in an email I sent to CNN, which is reproduced in Appendix III of this report. I cc'd this email to all the members of the Caltech/MIT Voting Technology Project. I received an email reply from one member, Charles Stewart, on 11/19/2004. Stewart seems to confirm that there is now no way to obtain the data used in the report:

We are aware that the data keep changing. We are also aware that others have dug up other exit poll estimates that precede the ones that we relied on.

POSSIBLE EXPLANATION FOR DATA CORRUPTION

There are indications, as I will explain, that the VTP group may have inadvertently used exit poll data artificially adjusted to make it match the official election returns. If that turns out to be true, that could help explain why the exit poll data match the official election returns better than fundamental statistics allow.

⁴ Indeed, the text of the report says that only 3 of the 51 exit polls differ significantly from the election returns. Assuming the authors mean that 3 of the 51 differ *at the one sigma level*, we can re-run the Monte Carlo simulation to check the odds that 48 or more of the 51 random numbers end up clustered within one sigma of the mean. The simulation shows the odds of such a clustering are less than 1 in 50,000.

It is standard practice for exit pollsters to present data that is adjusted to reconcile it with official election returns, once the election returns become available. Such adjustments may be at least part of the reason the exit poll numbers available from CNN.com today do not match the exit poll numbers the VTP researchers obtained earlier from CNN.com⁵ (see Appendix III).

Today, now that the VTP report is in circulation, the VTP researchers seem to understand that the exit poll data posted on CNN.com changed with time to make it match the actual election returns⁶. But the VTP researchers obviously did not understand this when they downloaded their data and wrote their report, since footnote 2 of their report says that people can reproduce the data anytime by downloading it from the CNN link. It seems that the VTP researchers may have inadvertently used artificial exit poll data.

If the VTP report used exit poll data that had been adjusted to match the election returns, then why is there any variation at all between VTP's exit poll data and the election returns? In other words, why don't all the data points in all three plots in the VTP report show a value of 0%?

I can think of at least three explanations.

(1) Roundoff error. The VTP researchers noted that the CNN website where they obtained their data does not provide the exit poll information directly in the form they need. Rather, the VTP researchers had to infer the data from other exit poll data, as described in their footnote 2. The numbers that CNN provided as inputs to this calculation had been rounded off to the nearest integer. Thus, even if the exit poll data had been adjusted to match exactly (to full precision) the election return data, the rounding process would add noise that would appear in the three VTP plots.

(2) Small discrepancies over election return data. Footnote 2 of the VTP report says, in part:

The primary election return data source for this paper is uselectionatlas.com, supplemented by official state web sites, to update the election returns.

(The report's exit poll data comes from one source, CNN, and the report's election return data come from other sources.) First note there is a possible typo in the footnote I just quoted. The website uselectionatlas.com does not exist, as of 11/21/2004 at 9pm PST. The authors may have meant uselectionatlas.org. There are some minor discrepancies between the election return data posted on uselectionatlas.org and those posted on CNN.com. For example,

⁵ This may also explain the discrepancy between (1) what the VTP report says CNN.com reported as the final exit poll support for Bush (49.8%) and (2) what Washington Post Managing Editor Steve Coll, as cited in VTP's report, said was the final exit poll support for Bush (48%).

⁶ In an 11/23/2004 email to me, cc'd to the rest of the VTP group, Charles Stewart wrote, "The exit polls were manifestly changed. I used whatever CNN had posted as of the morning after Election Day, probably around 9am, plus or minus. That's a moving target." Later in the same email he wrote, "It wouldn't surprise me at all that the exit pollsters re-weight the poll results to reflect the actual electorate that showed up on election day."

uselectionatlas.org says⁷ Bush received 61.07% of the vote in Alaska, while CNN.com says⁸ this number is 62%. Perhaps some minor confusion over the actual election returns contributed to the noise that appears in the three VTP plots.

- (3) It is possible the VTP researchers downloaded some of their exit poll data from CNN before the exit poll data was adjusted, and some after. I put forward this possibility in the email reproduced in Appendix III. When VTP's Charles Stewart replied via email on 11/19/2004, he did not specifically address this issue. He wrote only that the exit poll data are "a research product whose precise methods are unknown to us."

OTHER PROBLEMS WITH THE VTP REPORT

- The report seems to contradict itself on whether the national exit poll data agree with the election returns.

First the report says:

The first question to ask is, "How badly did the exit polls predict the outcome of the election?" The answer is, "not too badly."

Then the report says the difference between the exit polls and the election returns is "well outside the margin of error."

- The VTP report makes an assumption that limits the scope of its investigation, then claims too broadly that "there is no evidence, based on exit polls, that electronic voting machines were used to steal the 2004 election for President Bush."

The report limits the scope of its investigation by assuming the following:

If nefarious vote stealers had commandeered electronic voting machines on George Bush's behalf, then we would expect for the greatest discrepancies between the exit polls and the official counts to have been in the states that used electronic machines the most.

The report is implicitly assuming that all the electronic voting machines, regardless even of vendor, are programmed the same way. However, there is no reason for this assumption to be true. This assumption puts a strong limit on the scope of the VTP investigation, and thus it is inappropriate for the VTP report to make such a sweeping conclusion as it does.

⁷ <http://uselectionatlas.org/USPRESIDENT/data.php?&year=2004&datatype=national&def=1&f=0>, accessed 11/22/2004 at 5:15am PST.

⁸ <http://www.cnn.com/ELECTION/2004/pages/results/states/AK/>, accessed 11/22/2004 at 5:15am PST.

- The VTP report makes a statement that seems to dismiss statistics theory:
Even when they work well, exit polls are too imprecise to lay against the official count, unless every voter is included in the exit poll.
- The VTP report concludes with a quote from Steve Coll, Managing Editor at the Washington Post, in which he suggests that the only way the 2004 presidential election could have been stolen away from Kerry is if:
... a vast conspiracy carried out by scores and scores of county and state election officials was successfully carried off to distort millions of American votes.

I feel this quote illustrates a misunderstanding of how susceptible electronic voting machines are to rigging, and by whom.

David Dill, a computer science professor at Stanford, put it this way⁹:

If I was a programmer at one of these [voting machine] companies and I wanted to steal an election, it would be very easy. I could put something in the software that would be impossible for people to detect, and it would change the votes from one party to another. And you could do it so it's not going to show up statistically as an anomaly.

As Dill notes, the election could be rigged by a single individual working within one of the voting machine companies, without the help of “scores and scores of county and state election officials.” In fact, as is noted in the same New York Times article, county and state election officials are often enjoined from inspecting the software used by these voting machines:

Ms. Mercuri ran up against this last year, when she served as a consultant in a contested city council election in Boca Raton, Fla. Her request to look at the software inside the city's machines, made by Sequoia, to see if there were any bugs or malfunctions, was denied by a judge on the grounds that the technology was protected by trade-secret clauses. Sequoia, ES&S and Diebold routinely include such clauses in their contracts.

A misunderstanding of how electronic voting machines can be rigged may have compromised the approach VTP took in its search for statistical evidence for or against such rigging.

CONCLUSION

There is strong statistical evidence that the data used in the VTP report is corrupt. The data is corrupt in a way that removes the patterns the researchers were looking for.

In addition, even if the data were not corrupt, the conclusions given in the VTP report are worded too strongly, given the analysis performed in the report.

⁹ Excerpt from the New York Times, “Machine Politics in the Digital Age,” Nov. 9, 2003, by Melanie Warner.

APPENDIX I – CNN ELECTION DATA

This data was downloaded from <http://www.cnn.com/ELECTION/2004/pages/results/index.html> over the period Nov. 18-22, 2004. The Bush/Kerry numbers are the election returns. The last column shows the sample size for the exit poll.

	BUSH	KERRY	SAMPLESIZE
AL	63	37	736
AK	62	35	1177
AZ	55	44	1907
AR	54	45	1459
CA	44	55	2390
CO	52	47	2534
CT	44	54	872
DC	9	90	795
DE	46	53	772
FL	52	47	2862
GA	58	41	1618
HI	45	54	622
ID	68	30	801
IL	45	55	1434
IN	60	39	941
IA	50	49	2512
KS	62	37	667
KY	60	40	1050
LA	57	42	1683
ME	45	53	1991
MD	43	56	1065
MA	37	62	889
MI	48	51	2555
MN	48	51	2190
MS	60	40	799
MO	54	46	2264
MT	59	39	650
NE	67	32	785
NV	51	48	2189
NH	49	50	1883
NJ	46	53	1520
NM	50	49	2006
NY	40	58	1452
NC	56	44	2167
ND	63	36	687
OH	51	49	2020
OK	66	34	1577
OR	48	52	1064
PA	49	51	2107
RI	39	60	809
SC	58	41	1782
SD	60	39	1550
TN	57	43	1783
TX	61	38	1794
UT	71	27	816
VT	39	59	698
VA	54	45	1431
WA	46	53	2178
WV	56	43	1728
WI	49	50	2321
WY	69	29	761

APPENDIX II – MATLAB CODE FOR MONTE CARLO SIMULATION

```
n_sim = 300000;      % Number of simulations to run.
n_exit_polls = 51;  % Number of exit polls.
threshold = 45;     % Threshold for the number of polls less than one sigma from the
                   % true mean.

Y = randn( [ n_exit_polls, n_sim ] );

randn( 'state', 0 );    % So we get the same answer each time...

ct = 0;
for i = 1: n_sim
    %hist( Y(:,i) )
    %stem( Y(:,i), ones(n_exit_polls) )
    ff = find( abs(Y(:,i)) <= 1 );
    count = length( ff );
    if count >= threshold
        ct = ct + 1;
    end
end

disp( sprintf( 'ct = %d', ct ) )
```

APPENDIX III – EMAIL FROM ME TO CNN AND CC'D TO THE VTP GROUP

As shown, I originally sent this email to election.help@cnn.com. However, that address no longer exists, so a few minutes later I submitted this same text to CNN via a web submission form at: <http://www.cnn.com/feedback/forms/form2a.html?1>.

From: Leonard Wayne <lrwayne@earthlink.net>
To: election.help@cnn.com
Cc: nlo@caltech.edu, rma@hss.caltech.edu, bruck@paradise.caltech.edu, jkatz@caltech.edu, drk@caltech.edu, cmvest@MIT.EDU, selker@media.mit.edu, sda@MIT.EDU, berinsky@mit.edu, devadas@mit.edu, sgraves@mit.edu, rivest@mit.edu, cstewart@mit.edu
Subject: Caltech/MIT Voting Technology Project
Date: Nov 19, 2004 4:49 AM
Dear CNN:

There is a report by the Caltech/MIT Voting Technology Project that is widely circulating around the Internet and which purports to use data from cnn.com as its basis for analysis. However, the data in the report does not match the data that is currently on the cnn.com website (as of 11/19/2004, 2:30am PST). Do you have any insight into this? This concerns the controversy over the 2004 election and whether the new electronic voting machines in use might have been rigged. The Caltech/MIT report is titled, "Voting Machines and the Underestimate of the Bush Vote (Version 2, November 11, 2004)," and is available at:

<http://www.vote.caltech.edu/Reports/VotingMachines3.pdf>

Footnote 5 says:

Rhode Island gave 47.4% support to Kerry [typo - should say Bush] in the exit poll, compared to the actual 38.9%. The two other statistically [sic] differences were Oklahoma (59.2% exit poll vs. 65.6% official return) and New York (31.9% exit poll vs. 40.5% official return).

However, using the methodology described in footnote 2 of the report to compute the exit poll numbers based on the data on cnn.com at 2:30am PST on 11/19/2004, the computed exit poll numbers differ significantly from those listed in footnote 5:

Rhode Island: 47.4% (footnote 5) vs. 38.9% (11/19/2004, 2:30am PST)
Oklahoma: 59.2% (footnote 5) vs. 65.4% (11/19/2004, 2:30am PST)
New York: 31.9% (footnote 5) vs. 40.9% (11/19/2004, 2:30am PST)

I know that sometimes exit poll data reported by you and others is adjusted based on the actual precinct returns recorded by elections officials. Is it possible that the exit poll data on your website has changed over time for this reason? If so, is it possible that the authors of this report that is circulating so widely grabbed some of their data from the cnn.com website before the exit poll data was adjusted and some after the exit poll data was adjusted?

This last question is especially important, because if the authors used data that had been adjusted based on actual precinct data, then the central conclusions of the Caltech/MIT report may be in error.

Thank you in advance.

Leonard Wayne
3059 Sycamore Ave.
La Crescenta, CA 91214-3741
(818)393-5481 (W)
(818)957-4292 (H)

For reference, here is how I computed the exit poll data using the methodology described in footnote 2:

RI:

CNN: <http://www.cnn.com/ELECTION/2004/pages/results/states/RI/>

Accessed 11/19/2004 2:30am PST.

With 100% of the precincts reporting, Kerry gets 247,407 (60%) and Bush gets 161,654 (39%).

Bush: $.47*.41+.53*.37 = 0.3888$

Kerry: $.47*.57+.53*.62 = 0.5965$

OK:

CNN: <http://www.cnn.com/ELECTION/2004/pages/results/states/OK/>

Accessed 11/19/2004 2:30am PST.

With 100% of the precincts reporting, Kerry gets 504,077 (34%) and Bush gets 959,655 (66%).

Bush: $.48*.67+.52*.64 = 0.6544$

Kerry: $.48*.33+.52*.36 = 0.3456$

NY:

CNN: <http://www.cnn.com/ELECTION/2004/pages/results/states/NY/>

Accessed 11/19/2004 2:30am PST.

With 99% of the precincts reporting, Kerry gets 3,986,172 (58%) and Bush gets 2,793,745 (40%).

Bush: $.45*.42+.55*.40 = 0.4090$

Kerry: $.45*.56+.55*.60 = 0.5820$