

# **SOURCES OF VARIATION IN PRE-ELECTION POLLS: A PRIMER**

Or

*Why Different Election Polls Sometimes Have Different Results*

**Cliff Zukin,**

Professor of Public Policy and Political Science

Bloustein School of Planning and Policy & Eagleton Institute of Politics,

Rutgers University

October 24, 2012

## ***Preface***

*During any campaign season, a multitude of polls about the candidates' relative positions appear in the broadcast and print media. AAPOR representatives are called on repeatedly to explain the how and why of these surveys – especially why their results in the horserace might differ one from another. Cliff Zukin, former AAPOR President, graciously agreed to revise and update the following primer to help us answer inquiries about these polls, and AAPOR thanks him for his excellent attention to this task. The result of this work is a short primer which goes a long way in explaining election polling to journalists and others. While some readers may differ with one or another element of this essay -- which is not meant to represent a formal AAPOR position -- it gives journalists and others the background to make sense of what can be perplexing differences in a complex and evolving field.<sup>i</sup>*

*The primer addresses the surveys conducted by and for the media – not surveys conducted by political pollsters to assist their candidates in conducting campaigns. Those campaign-sponsored polls may use many but not precisely the same practices and methods as the published polls described here. – Paul J. Lavrakas, 2012-2013 President, AAPOR.*

## **A Primer**

Election polls are a special breed among public opinion surveys. They call for more judgments—the art rather than science of the craft—on the part of the pollster than other types of polls. And this brings into play a host of other reasons why the estimates of well established and well done pre-election polls may differ from one another, even when these polls are conducted at a similar point in time. This primer is meant to be a guide for journalists, academics, and anyone closely following polls in the 2012 election.

## Real Sampling Error

Most seasoned political observers are familiar with the notion of *sampling error* in public opinion polling. That is, because we select a sample to represent a population we are making an *estimate*, of candidate preference for example, and making an inference from our sample back to population. This margin of error, expressed as “plus or minus” a number of percentage points, is the most commonly known source of variation for why polls may differ. In this election year we often hear statements, such as Obama leads Romney by three points, 47% to 44%, with a sampling error of plus or minus three percentage points.

What is less commonly known is that the margin of sampling error does not apply to the spread between the two candidates, but to the percentage point estimates themselves. If applied to the three point spread the three point margin of error would seem to say that Obama’s lead might be as large as six ( $3 + 3$ ), or as little as zero ( $3 - 3$ ). But when correctly applied to the percentage point estimates for the candidates Obama’s support could be between 50 and 44 % ( $47 \pm 3$ ), and Romney’s between 41 and 47 % ( $44 \pm 3$ ). Thus the range between the candidates could be from Obama having a 9 point lead ( $50 - 41$ ) to Romney having a 3 point advantage ( $44 - 47$ ). So, sampling error is generally much larger than it may seem, and is one of the major reasons why polls may differ, even when conducted around the same time.

## Sampling Error

- Sampling Error is a theoretical minimum
- It is only one kind of error, but it is quantifiable
- It applies **not to the GAP** between candidates, but to each point estimate

MOE  $\pm 3$

Obama 47	Romney 44
Obama $47 \pm 3$	Romney $44 \pm 3$
Obama 44 to 50	Romney 41 to 47

18

## Sampling and Coverage Concerns: Household Selection

Sampling is the foundation of scientific survey research, and is based on the branch of mathematics having to do with probability theory. In short, a probability sample is necessary for its numbers to be legitimately generalized from the sample back to the population from which it was drawn. This assumption is not warranted in the case of non-probability samples. Any poll comprised of self-selected respondents, including call-in polls and Internet or Web surveys where people volunteer to participate in response to an open invitation are **non-probability** samples. It is meaningless to calculate a sampling error in a non-probability sample, it forfeits any claim of generalizability, and it may be a disservice to the public to report the results of such pseudo-scientific surveys.

NY Times: *In order to represent the population statistically, a survey should be based on a probability sample.*

-ABC/Washington Post: *Methodologically, in all or nearly all cases we require a probability sample, with high levels of coverage of a credible sampling frame. Self-selected or so-called "convenience" samples, including internet, e-mail, "blast fax," call-in, street intercept, and non-probability mail-in samples do not meet our standards for validity and reliability, and we recommend against reporting them*

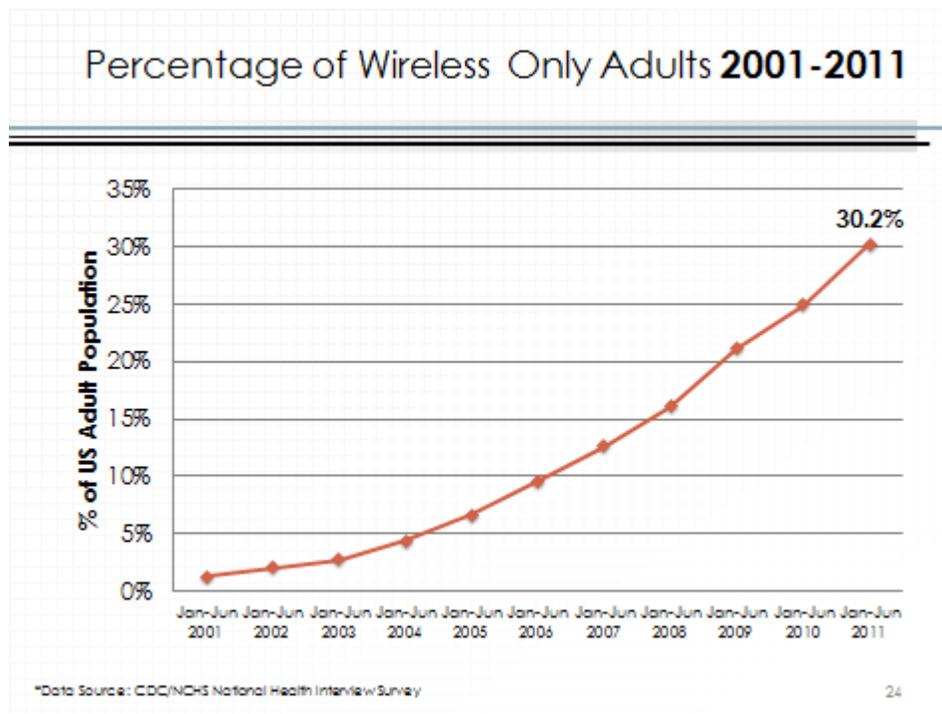
Most pre-election surveys are conducted by telephone, using one of two types of sampling frames, or definitions of who is eligible to participate. The most common approach in the US is what is called a **RDD** sample, short for random digit dialing. In this case samples of (hopefully *both* landline and cell phone) telephone area codes and exchanges are taken, and then random digits added to the end to create 10 digit phone numbers. The first step ensures proper distribution of phone numbers by geography; the final step, adding the random numbers, makes sure that even unlisted numbers are included. In the case of the landline RDD sample, a randomly designated respondent is then asked to participate in the survey. This is the standard practiced by most public pollsters.<sup>ii</sup>

An alternative is called registration based sampling, or **RBS**, and this is the method used by most commercial pollsters working for political candidates and parties. This begins with a sample of individuals drawn from publicly accessible lists of registered voters, to which phone numbers are then matched. This is less costly and more efficient, as almost all calls result in reaching a working phone number and a registered voter, which is not

true of a RDD sample. The primary disadvantages of RBS surveys is that they miss people who have recently moved or are *newly registered to vote*, which may be a non-trivial portion of the electorate in some states, or the country as a whole, and that it is largely confined to **listed** telephone numbers,<sup>iii</sup> meaning a substantial portion of the electorate could be missed. Also, the purging and updating of voter registration lists historically has varied widely from state to state, so the accuracy of RBS sampling will vary, although official state lists have become more consistent with funds from the Help America Vote Act.

## Cell Phones

In 2004 I wrote, “Although weighting may be used to try to make up for this shortfall, (see “Weighting” below) ‘cell phone only’ households are a relatively new phenomenon and we don’t yet fully understand the consequences of this bias. In 2012, we know.<sup>iv</sup> The distribution of various types of telephones in the US has changed so much since the 2004 election that I assert it is not possible to conduct a meaningful study of the general public or a representative sample of voters through a landline-only survey.



Consider the following:

- The percentage of wireless adults, or **cell phone Only**, was 4.4 percent in 2004; It increased dramatically between 2004 and 2008, to just over 15% at the time of the election. But it has exploded since then, essentially doubling that number to 32% by the end of 2011.<sup>v</sup>
- Adding those who are **cell phone Mainly**, where at least half of incoming calls in a household are answered on cell phones, to the cell phone only group, the total that might be missed in a **landline only** phone poll may be **half** of the public. A very conservative seat of the pants estimate is that this is at least one-third of those likely to vote on November 6.
- In 2012, the American public has four times as many *cell phone only* users (32%) than *landline only* users (8%).
- If a survey organization conducted a random sample of 1,000 Americans by landline only, 18 to 29 year olds *should* make up 22 percent of the sample; instead they will make up about six percent.<sup>vi</sup>

***Therefore, a critical question to ascertain from polling organizations is whether or not they have included cell-phone only respondents.*** Most of the better survey organizations will have at least 300 cell phone only respondents in a sample of about 1,000 respondents, which allows them enough to reliably weight the two sampling frames (individual and household) after interviewing and before analysis. Phone status interacts with how the interview was conducted here. Most RDD landline surveys use “auto dialers” to reach respondents to eliminate the need for interviewers to hand-dial the numbers. ***Federal law currently prohibits using an auto dialer for cell phone calls.*** That means that interviewers must hand-dial thousands of cell phone numbers, increasing field costs by a significant amount.

Some pre-election telephone polls use no live interviewers at all and rely exclusively on recorded voices. These are called ***IVR polls***, for *interactive voice response*, or sometimes robo-polls in the vernacular. A few organizations using IVR hand dial cell phones and add them to their landline samples, but more do not because of the cost. ***If they do not, be very careful reporting the findings of these polls that cover only one half of the American public.***<sup>vii</sup> The evidence shows that polls without cell phones may underestimate the strength of Democratic candidates.<sup>viii</sup>

## A Growing Difference Between Landline and Dual Frame Samples

	Landline and cell sample	Landline sample	Diff
<b>2010 Congressional vote</b>	%	%	
Republican	49.6	52.1	2.5
Democrat	42.0	39.4	2.6
Other/Don't know	<u>8.4</u>	<u>8.5</u>	
	100	100	
<b>Rep-Dem margin</b>	<b>+7.6</b>	<b>+12.7</b>	<b>5.1</b>
<b>2008 presidential vote</b>			
McCain	41.6	42.7	1.1
Obama	49.8	48.5	1.3
Other/Don't know	<u>8.6</u>	<u>8.8</u>	
	100	100	
<b>Rep-Dem margin</b>	<b>-8.2</b>	<b>-5.8</b>	<b>2.4</b>

Source: Pew Research Center

### Respondent Selection

Sampling a household is only the first stage of landline RDD surveys, and in itself is insufficient to ensure a representative sample. If interviewers simply spoke with whoever answered the phone the resulting sample would be older and more female than the population as a whole. Given who answers the phone when the landline rings, we would certainly expect IVR polls to be disproportionately composed of older women, who, one would suspect, have a different issue agenda than a cross section of the public as a whole, and a particularized response to the two candidates as well. To produce a representative sample, survey organizations must also go through a respondent selection procedure

among potential eligible respondents *within* the household. Some organizations use a random technique, such as the “last birthday” technique, where the interviewer asks to speak to whoever had the last birthday in the household. There are other techniques of randomization, but the idea is to ensure that everyone has an equal, or at least *known*, chance for inclusion. Other organizations use a systematic technique, such as asking for the youngest male/oldest female at home, that have produced empirically representative samples in the past.

Some surveys, including almost all IVRs, go through no respondent selection process in the household. Whoever answers the phone may be told: touch 1 if you plan to vote for Romney, 2 if voting for Obama). Such polls use no selection technique when contacting a household, but instead try to compensate for this by weighting after the data were collected and before analysis. Like non-probability samples, these types of non-probability respondent selection mechanisms compromise a fundamental tenet of probability sampling and should require an accounting and justification before being reported.<sup>ix</sup>

## **Timing and Field Procedures**

Timing of course refers to when a poll is done. And as all pollsters are fond of saying, even the best pre-election poll is a no more than a snapshot in time. Polls don’t predict; they describe the situation of the moment. Obviously, pre-election polls with different field dates may yield different results, as voter preferences may change with time. However, a largely invisible reason for differences is that polling organizations have different field procedures. Field procedures refer to the ground rules under which the interviewing is done. And, there are some tradeoffs to be made.

For example, a field period of seven days would allow for a number of callback attempts to reach the selected respondent before allowing substitution with a new household and respondent. But campaign events may happen in those seven days, making that poll harder to interpret. A three day poll may focus more narrowly on a particular point in time, but likely at the sacrifice of callback attempts. Callbacks matter because one respondent may not be the same as the next; extra field time may be necessary to reach younger voters for example, who may be more Democratic in orientation than others. So factors like the number of callbacks, days of interviewing, and response rates may also be reasons why polls purporting to measure the same thing give different results.

*Tracking polls*, or polls where interviews are conducted every day either released on their own or aggregated with others to some consistent base, such as “the last 3 days,” are a special case. Given that their callback procedure to reach their primary respondents is compromised by the length of time in the field, they may be more useful for spotting trends of voters moving up or down (reliability), and somewhat weaker at estimating vote choice (validity).

## Question Ordering and Wording

It has long been known that the ordering and wording of questions in a survey can affect the results. In ordering, responses to questions asked early in the interview schedule may affect later ones, as frames of reference are set, and respondents strain to be consistent in their responses to interviewers. For example, a survey that asked respondents a set of questions on the economy before asking for whom they planned to vote could lead to a bias in favor of Romney, who is currently perceived as stronger on economic issues. And a line of questioning that asked people to assess how the government did in dealing with Osama bin Laden could lead to a bias in favor of Obama if the “vote” question was asked after this line of questioning.

In order to minimize this problem, most researchers will ask the horserace question (*If the election were held today, for whom would you vote...*) before any other substantive election question on the survey. This does not include neutral questions about whether people are registered, or how interested they are. After all, when people go into the voting booth on November 6 they will have had no warm up questioning on issues or candidate qualities. However, perhaps in hopes of simulating the campaign, some polling organizations begin their surveys with substantive policy or election-related questions before asking about vote intentions. When interpreting poll results it is always useful to know the context in which a question was asked. While two polls may have asked the horserace question in the same form, one may have done so after unconsciously pushing some respondents in one direction or the other by earlier questioning. So question ordering also becomes a source of possible variation in the results among published polls. Best practice is to ask questions that might influence vote choice **after** the horserace question.

The wording of questions—even the horserace question—may also vary from one poll to the next. Some polls will ask a two way vote intention question, naming only the major party candidates but recording all answers, while others will explicitly add a third party candidate’s name. (For EG: Ralph Nader in 2000, not an issue in the 2012 presidential, but may be in state races, or ask about the Green party, or add a response choice of voting for an independent candidate.) Some polls ask the horserace question twice in the same survey, once with the two major party candidates and once with a more expansive list of candidates. But one of these must be asked before the other, and the order may influence responses. While most polling organizations asking about the candidates add their party labels as a cue, some may just name the candidates. And trial heat questions that also name the vice presidential candidates may produce somewhat different results than when only the presidential candidates are mentioned. There is even some evidence that there is a slight bias in favor of whichever candidate is named first in survey responses, so some organizations **rotate** the names of the two candidates while other do not. So differences in question wording may also be a reason why polls have small differences in their reported findings.

## Weighting

Weighting is an important and common practice in survey research. Even the best polls cannot interview a perfect sample, due to non-response and non-coverage, among a variety of reasons. (Non-response occurs when people who are sampled refuse to take part in the survey or are never contacted during the field period; non-coverage occurs when not all people who will be voting are included in the sampling frame—an Internet survey would miss voters who do not have access to or use the Internet; a telephone survey would miss those without any phone service, for example.) Thanks to the U.S. Census we know how many people in the entire U.S. have a few fixed characteristics, such as age, education or race/ethnicity. When we look at who we actually interview in our samplings, we can adjust—or weight—for these characteristics to make sure they are correctly represented. For example, if we knew that **30 percent** of the adult **population** had graduated from a four year college, and we had 45 percent in our **sample** of 1,000 report they had graduated from a four year college, we would need to weight these respondents by a factor of .67 to make sure the data reflect the correct proportion of this group. [The math here is straightforward: Take the 45 percent of college graduates, count each as .67 of a person, and they will contribute to the pool of all answers as if they were 30 percent of the total. (.45 x .67 = .30.)]

As ubiquitous and powerful as it is, it is important to note two limitations of weighting when talking about election polling: the only population parameters researchers can have confidence weighing to must be (1) known and (2) stable. If there isn't a fixed known parameter in the electorate, such as the percentage of the electorate with bumper stickers, it can't be weighted to. And obviously, one cannot weight to an event that has not yet happened, such as the percentage of Election Day voters earning over \$100,000. There are two controversial aspects of weighting as it applies to election polling—weighting by those most likely to vote in the election, and weighting by party identification.

**Weighting to the demographic compensation of the electorate:** Because of the Census, we know the characteristics of the general adult population. But of course **we do not know what the voting public will look like on Election Day**. We don't know how many of any racial, educational or generational group will be going to the polls until after the fact. In the 2000 election, the national exit poll estimated that African Americans made up about 10% of the electorate, and about 90% of these votes went to Al Gore. What if 7% of a polling organization's sample is made up of African Americans in 2012? What if it is 13%? It will obviously make a difference in the horserace estimate, but we won't know which is correct until Election Day. And, of course, the past is no guarantee of the future. So, the pollster's dilemma is "What do we weight to?" Most pollsters of published surveys first ask a sample of the general population about their race, gender, age, etc., and then weight their data to what a random sample of the population should look like, and then go on to pull likely and non-likely voters out of that big (already weighted) general population sample. Some, however, weight to a picture of what they *believe* turnout will be, based on past experience and elections -- and not everyone doing

so is painting the same portrait. (This practice is more common among campaign-sponsored polls than published polls.)

**Weighting party identification:** A second issue concerns whether polls should be weighted to reflect an assumed distribution of the electorate by the political party of respondents. A party identification question, generally placed near the end of the survey, asks people to state whether they consider themselves to be a Democrat, Republican, independent or something else. The vast majority of pollsters do **not** feel it is appropriate to weight by party. The scholarly literature comes down firmly on the side that party is not a fixed attribute, like race or gender or age. It is an attitude, and peoples' responses to this question change based on circumstances and events. And indeed, the American public does show fluctuation in partisanship over time, as well as individual changes. A small number of pollsters do weight by party, but that is tantamount to guessing what the electorate will look like on Election Day, which of course is unknown.

Party ID is the most critical variable predicting the vote. By November, it's highly likely that at least 85 percent of Republicans will vote for Romney and at least 85 percent of Democrats for Obama. A two or three point difference in the estimation of the partisan makeup of the electorate will easily lead to polls that differ by two or three points, all other things being equal. And, the empirical evidence to date in 2012 is that there are differences in the estimated partisan characterization of the electorate, based on the type of survey conducted. Mark Blumenthal, blogging on polls at the *Huffington Post*, reported a 13 percentage point variation in the percentage Democratic as estimated by Internet and telephone polls this past summer.

## The Special Case of Party ID

Party ID	Phone	IVR	Internet
Democrat	33%	36%	46%
Independent	35%	29%	15%
Republican	30%	34%	36%
Totals	98%	99%	97%

\* Mark Blumenthal @ Huffington Post August 24, 2012

76

## Likely (Probable) Voters

Another problem all pollsters face every election is the *over-report of the intention to vote*. When respondents' self-report of intentions in pre-election polls have been compared to actual turnout (again, known only after the election) we have historically found a large over-report of voting intentions. So the pollsters' dilemma here is to separate the wheat from the chaff: Of all those saying they will vote on Election Day, which ones will really do it, and which ones will stay home? And, of course, people change in their commitment to voting as the campaign unfolds. Respondents are probably better able to tell if they really are going to vote as it gets closer to Election Day. This means that the definition of a likely voter is somewhat of a moving target, compared to the definition of registered voters, for example.

Research finds no magic bullet question or set of questions that can reliably determine likely voters with 100 % accuracy. Thus, different organizations have different ways of estimating who are probable voters. Most polls ask a combination of questions that cover three areas that are highly correlated with voting: a) self-reported vote intention; b) measures of engagement (following the election closely, interest, care who wins); and c) past voting behavior (voted in prior elections). They then combine responses to create an index that gives each respondent a total score. Most then use a cutoff point so that only the candidate preferences of the "most" likely voters are used, and the choices of others are discarded. But even while most use such a scale, the component questions that go into the scale differ, and so this too is a source of variation among polling organizations.

There are other approaches as well. A more efficient but perhaps more risky strategy uses a single question or two of reported intention and does not complete the interview with those not passing the screen. For example, a poll might ask about someone's chances of voting on a scale of 0 to 10, and only continue interviews with those who gave themselves a 10. This could result in only 40 percent passing this screen. If real turnout on Election Day was 56 percent, the underestimation might be biased if Democrats were less likely to be among the initial 40 percent but overrepresented in the next 16 percent slice of the electorate. While most polling organizations use a cutoff point for likely voters (take all of those in the top 56 percent and none of those in the bottom 44 percent), others may give voters weights based on the probability of voting to everyone in the sample rather than using a cutoff. And still others may simply use a fixed set of screening questions that have worked well for them in the past, leaving a lot of room for variation in the vote choice estimates produced by different polls.

A second issue in determining likely voters is estimating *how many* there will be, which may affect the division of the vote. In the last three presidential elections the percentage of the voting eligible (not registered) public turning out to vote varied by almost seven percentage points (2000, 54.2%; 2008 61.1%). Suppose a choice of a cutoff point of 61 percent gives an estimate that Obama leads Romney by seven percentage points. But

when pruning the expected electorate to 53 percent, it may be that the data show Obama leading by just five points (See box below). So, another source of possible differences is what percentage of voters is let in during the likely voters scoring process. Moreover, some may start with a base of registered voters (72% in New Jersey in 2000) while others may work with a percentage of the voting age population (52%) as their base. There are also differences in the voting age and voting eligible populations in each of the states. Thus while all polling organizations will release figures for who they believe are *likely* voters, no two organization will define them in exactly the same way. It is worth noting that estimates of likely voters generally come into congruence as the election gets closer.

Vote Choice	Top 61%	Top 54%
Obama	50	53
Romney	43	48
Undec.	7	4
<b>Margin</b>	<b>7</b>	<b>5</b>

### In Summary

There are a number of choices to be made in the course of conducting election polling beyond sampling error. We call these “house differences” where different organizations have differ ways of doing this type of research. To look for trends it is probably safest to compare polls done by the same organization at different times, rather than to try to compare polls with different methodologies done at similar times. Given the unique nature of election polling, it is likely that outsiders may look at them with puzzlement and ask “What’s going on?” I hope this essay is helpful to our journalistic and other colleagues in understanding some of the sources of variation in election polling. From the inside, those of us conducting election polls see a fair amount of consistency in findings amid the complexity of a science-based-art.

## Why Election Polls May Vary by a Few Percentage Points

- Sampling error
- Length of field period
- Live interviewers vs. IVR
- Type of sample used
- Mode of administration
- Respondent selection
- Likely voter indices
- Question wording & ordering
- Weighting

**Actually, it's a wonder they are as close as they are!**

77

---

<sup>i</sup> The ideas and views contained in this document are those of the author. This is an update of a piece I wrote in 2004. Thanks to Peyton Markley Craighill, Krista Jenkins, Paul Lavrakas, Diane Colasanto, Rob Daves and Mark Szeltner for reading and commenting on earlier drafts of this paper.

<sup>ii</sup> There has been an increasing use of a new sampling frame for probability-based sampling known as **ABS**, for Address-Based Sampling, since the last presidential election. Researchers have access to the master address file of the USPS, through a licensed vendor. A random sample of respondents is pulled from this frame. Names and addresses are run through various data bases to match a phone number with name and address. A possible weakness with this method is in the efficiency of matching numbers to names. This form of sampling is not extensively used in election polling (yet).

<sup>iii</sup> Some unlisted numbers are “findable” when put through various commercial data bases.

<sup>iv</sup> Hats off to AAPOR and the survey research industry for figuring out how to incorporate cell phones in dual frame designs and keep the sample survey valid.

<sup>v</sup> Data Source: CDC/NCHS National Health Interview Survey

<sup>vi</sup> Data Source: The Pew Research Center for the People and the Press

<sup>vii</sup> I’m not making the argument that IVR polls are bad per se. Not at all. I’d much rather use them on polls asking questions in sensitive areas, such as drug use and sexual behavior. But the tool needs to be fit to the research problem at hand.

<sup>viii</sup> Data Source: The Pew Research Center for the People and the Press

<sup>ix</sup> In surveying those reached via a cell phone number in the USA, and assuming the person who answers is an age-eligible adult, most pre-election pollsters proceed to interview the person who answers rather than determining if the cell phone is answered by more than one adult. This is the case because it is thought that few cell phones in the USA among the voting public are shared devices.