Man versus Machine: An Experiment Comparing Human Interviewers and Interactive Voice Recording Technology

Joshua D. Clinton[1]

John Lapinski[2]

Steven Rogers[3]

Abstract:

With the rising demand for instant feedback on public opinion, surveys using interactive voice recording (IVR) technology have become increasingly popular to assess public opinion on political issues. Despite research showing that survey mode significantly affects responses to certain types of questions, we know little about how these so-called "robo-polls" differ from surveys conducted personally by human interviewers. To rectify this lack of knowledge, we field an experiment to compare a human interview poll, a one-day IVR poll, and a multi-day IVR poll using the same sampling frame, the same questions, the same calling period, and the same weighting procedures. Relative to a landline human poll, we discover that landline IVR polls have lower response rates and substantially greater drop off during interviews. We additionally find that respondents who manage to complete an IVR survey are more likely to be older and female. Finally, and perhaps most importantly, we find that even after weighting to account for the demographic differences we detect, public opinion estimates from the two modes sometimes differ beyond what can be attributed to chance alone.

Word Count: 7281 (excluding references, tables, figures and appendices)

[1] Associate Professor, Department of Political Science Vanderbilt University. Josh.clinton@vanderbilt.edu

[2] Associate Professor, Department of Political Science, University of Pennsylvania. Director, Elections. NBC News.

[3] Assistant Professor, Saint Louis University. smrogers@slu.edu

The media and the American public have an almost insatiable appetite for public opinion polls dealing with politics and elections. At their best, polls provide us with an understanding of what the public believes and why they believe what they do. It can even be persuasively argued that well-crafted polls enhance certain aspects of democratic governance (Gallup and Rae 1940). However, whereas good polls can promote awareness of public opinion on important issues before the country, poorly designed and implemented polls may just as easily cause harm by mischaracterizing public opinion.[4]

What constitutes a "good" poll is therefore a critical debate that is subject to continual discussion and refinement. As available polling technologies and societal norms evolve, there is a constant pressure for conscientious pollsters to ensure that their methods are able to reach the target populations of interest (Blumenthal 2005; AAPOR 2009; DiCamillo 2010). Changes in polling technology can create new ways of reaching and interviewing respondents (Stern, Bilgen, and Dillman, 2014), but even with the large literature devoted to better understanding the ability of different polling technologies to accurately ascertain reality (e.g., Parry and Crossley, 1950; Presser and Stinson 1998; Belli et al., 1999; Tourangeau, Steiger, and Wilson 2001; Holbrook, Green, and Krosnick, 2003; Kreuter, Presser and Tourangeau 2008; Couper 2011; Lind et. al. 2013), the consequences some technological changes have on the measurement of public opinion are not always well understood. Some news organizations, such as The Washington Post, ABC News, and NBC News (Anand 2010; Moulitsas 2011; Cohen 2012), have created internal standards as to what polls should and should not be reported, but only by comparing the relative performance of various polling modes holding all else equal can we truly identify

[4] For example, Frank Luntz's work surrounding the Contract with America was found in violation of AAPOR's ethics code (AAPOR 1997)

whether the various polling technologies produce important differences in the estimate of public opinion.

One polling technology – the Interactive Voice Recording (IVR) poll – has been prominently featured by many media organizations. The steady supply of IVR polls and the incessant need for current polling data create what appears to be a perfect match. For example, in the 2012 Republican primary campaign more IVR polls were conducted and publicly reported than any other type of poll; following the New Hampshire primary, there were 106 IVR polls, but only 53 human polls conducted within 4 weeks of a state's primary election (Clinton and Rogers 2013).

One reason for the prominence of IVR polls is that they can be conducted quickly and relatively cheaply – it is possible to get hundreds of survey responses in a single night (Stern, Bilgen, and Dillman, 2014). During the 2012 Republican primary, for example, the average IVR poll had a field period of 1.57 days and sample of 816 respondents. In contrast, polls with human interviews took an average of 4.34 days to conduct, and they contained an average of only 529 respondents. The ability to deliver near-instant feedback on public opinion with a large enough sample to detect subtle shifts in public opinion creates a nearly irresistible combination. Together with the greater costs involved when using human interviewers, IVR polls at times provide the only assessment of public opinion in some states or election contests.

Despite their increasing prominence, it is not entirely clear how well self-completed IVR polls perform relative to a political poll personally conducted by human interviewers.[5] Some

[5] In analyses of non-political surveys, Kreuter, Presser and Tourangeau (2008) report an experiment in which a survey is conducted either by a computer-assisted telephone interview (CATI), IVR, or online. Asking questions of university alumni for which they had validation data, they find that respondents misreport in the socially desirable direction most in CATI administration and least in web surveys. Misreporting on IVR surveys falls between the two. In earlier work, Tourangeau, Steiger, and Wilson (2001) demonstrate consumer-satisfaction surveys receive more positive ratings when conducted via CATI than IVR, an effect the researchers attribute to the interaction with another human.

have raised questions about exactly what is being done to produce estimates using robo-polls (e.g., Cohn 2013a; Cohn 2013b; Elliot 2010), and others have compared the accuracy of IVR and human polls (e.g., AAPOR 2009; Silver 2012; Clinton and Rogers 2013). The inner-workings of IVR polls however largely remain a "black box." It is unclear how IVR and human polls compare across many dimensions that are typically thought critical to assessing the quality of a poll. This lack of information makes it difficult to judge the quality of IVR polls.

We seek to remedy this deficiency by reporting the results of an experiment we conducted using parallel surveys with human interviewers and IVR technology in the state of Tennessee. Using an identical state-level sampling frame and comparable field periods, we ask the same questions on three landline surveys: a survey with human interviewers, an IVR survey conducted in a single day, and an IVR-survey spread over multiple days. Our research design allows us to compare the response rates of each mode, who responds to each type of survey, how responses to identical questions vary, and the extent to which differences in opinion are mitigated by weighting procedures.

The results of our experiment are informative in several respects. In regards to data collection, we confirm that IVR response rates are lower than those of human surveys. Comparing the ratio of completed interviews to the sample of numbers called reveals that whereas 2.7% of numbers called by a human interviewer resulted in a completed interview, only 2.5% of identically selected numbers resulted in a completed interview using IVR technology. Moreover, only 61% of the respondents who start our three minute long IVR polls complete them. We additionally find that IVR technology exacerbates known issues with landline samples (Rosenthal 2010; Keeter 2014; Blumberg & Luke 2013). Men and younger people were less likely to respond to our landline IVR polls than a landline poll conducted by human interviewers.

Over 2,000 people completed our IVR surveys, but there were only 96 males under the age of 40 who did so, and more than 50% of the completed interviews were from individuals over the age of 64. Finally, even after weighting the sample of respondents, statistically distinguishable differences in estimated public opinion, such as partisan identification, were sometimes present.

Our results suggest that IVR polls may provide an assessment of public opinion that is similar to a human landline poll, but they may also produce estimates that differ by more than what can be attributed to random chance alone – particularly when attempting to measure the opinion of particular groups such as males or those who are under 40. There is suggestive evidence that systematic responses biases may be present, but it is difficult to determine whether bias is attributable to mode effects, selection effects, or perhaps some of both. A final important difference is that the high frequency of break-offs that occur among the general population we observe on our three-minute IVR poll suggests it may be very difficult to use this mode to probe public opinion in any depth.

Our paper proceeds as follows. In Section 1, we describe the research design we use to compare a human interview poll to single day and multi-day IVR polls. In Section 2, we characterize the response rates and drop off rates for the three polls we conduct, and we contrast the demographics of the respondents to each poll to Census estimates. Section 3, evaluates how the weighted and unweighted opinions of the three polls differ, and Section 4 concludes by discussing the larger implications of our study for understanding IVR polls.

1. Research Design & Methodology

To compare the impact of human interviews and IVR technology we conduct three simultaneous polls of randomly selected adults living in Tennessee. We specifically seek a better understanding of the extent to which polls interviewing the same general population and

conducted using the same questions and mode – telephone – produce different respondents and responses depending on whether the interviewer is conducted more personally by a live human being compared to a self-completed survey administered by a pre-recorded human voice (e.g. Bougher and Prior 2013). Our experiment complements the recent important research regarding how answers to telephone surveys compare with online (e.g., Christian, Dillman, and Smyth, 2008; Malhotra and Krosnick 2007; Ye, Fulton, and Tourangeau, 2011; Yeager et. al. 2011) or in-person interviews (e.g., Holbrook, Green, and Krosnick, 2003), but unlike prior analyses of IVR polls (Kreuter, Presser and Tourangeau 2008), we examine responses to a political survey where there is no verifiable "right" or "wrong" answer because these questions are often the focus of journalistic attention when covering elections and other political events.

In our experiment, one poll uses human interviewers, and two use IVR technology with varying field periods. We focus on a single state for our investigation because the increasing popularity of IVR polls is partially due to the fact they can cheaply and quickly poll areas that are infrequently surveyed using traditional survey methodologies. By focusing on a so-called "red state" where little polling is typically done, we study public opinion among a political constituency where IVR polls are most likely to provide the only assessment of public opinion. Moreover, the fact that we poll in a politically homogenous state may suggest that any differences we find in the measurement of public opinion is likely a lower bound – if we find differences in public opinion in a state that agrees on most political questions, we may suspect there to be even larger differences in a state where opinion is more divided.

To ensure that our comparison is as similar as possible, we use the same sampling frame of landline numbers for all three polls. The vendor conducting our landline human interviews (Princeton Survey Research Associates International) purchased three sets of numbers from the

same sampling frame from the same company (Survey Sampling International).[6] Because the numbers being called by human interviewers associated with Princeton Survey Research Associations International and the numbers being called by the IVR technology of Precision Polling are identically generated, any differences we detect between the polls cannot be attributable to differences in the sampling frame.[7]

A known limitation of IVR polls is that respondents are less likely to stay on the line to answer questions. As a result, our IVR polls are only able to ask a small subset of the questions we are able to ask on our 20-minute human survey. To maintain comparability, we therefore ask the same initial questions in our human interview survey and our IVR polls (see Appendix A for the IVR survey instrument). In addition to questions about the approval of various political leaders, we also asked about demographics and other politically relevant characteristics (e.g., registration status and partisan self-identification) at the end of the survey. Our IVR polls used the voice of a middle-aged white female.

All three polls attempted to contact numbers during the same time frame during the week -- 5:00 PM to 9:00 PM (accounting for the time zone of the number being called). Multiple callbacks were attempted for all three polls, but some replicates were not called the maximum number (5 times in the human poll and 3 times in the IVR polls).

Despite these similarities, some mode differences are inevitable. In addition to the limited length of our IVR polls, the human call center was limited in its ability to work numbers. To get 573 completed landline interviews, human interviews called 21,258 unique numbers between November 20 and December 5, 2013 (excluding holidays). In contrast, the one-day

[6] In particular, for each IVR poll, we purchased 50,000 numbers using a "Random A" sample type from active blocks with a minimum of 3 working blocks. The numbers that we purchased were unscreened for businesses or disconnects.

[7] Both PSRAI and Precision Polling were aware of the design of our research project and worked with us to ensure the integrity of our research design.

IVR poll conducted on November 20 called approximately 50,000 unique numbers - as well as callbacks two hours later for every number not reached – in a single night. [8]

IVR polls can reach more respondents in a single night than human interviewers, but it is unclear whether a single-day poll suffers from an inability to reach eligible respondents because of the limited ability to re-contact numbers. To explore this possibility, we conducted a second IVR poll using a multi-day field period. In total, we conducted two IVR polls – a one-day IVR poll conducted on November 20, 2013 with callbacks every two hours, and an IVR poll conducted from November 20 to November 24, 2013 with callbacks once a day up to 3 times in total.[9]

Table 1 presents the disposition of the numbers called in each of the three studies. In our human poll, we were able to use the trained interviewers to identify the disposition of the numbers being called. Whereas a human interviewer can determine whether a number is ineligible because it is a business or non-English speaking household, assigning disposition codes using IVR technology was more difficult. Unless a phone number was non-working, the IVR technology would recall the number repeatedly until a connection was made. Since we use the same sampling frame for telephone numbers in all three polls, the fact that the working number rate is so much lower in the human poll compared to the IVR poll in Table 1 – 19.2% versus 95.7 % for the one-day IVR poll – strongly suggests that there are many numbers that are classified as "valid" by the IVR software that are not. For example, 36,049 numbers were never

[8] For the IVR polls, we additionally attempted to identify whether the number was an answering machine when conducting the poll so that machines with answering machines could be called back. The technology used to detect answering machines is imperfect – it is essentially an algorithm that estimates whether a number is a human by waiting a certain amount of time for a response under the assumption that humans respond within a certain amount of time after answering a telephone call. However imperfect, this is necessary for calculating response rates because the IVR technology would otherwise count a message left on an answering machine as an "incomplete" rather than a number that needs to be recalled.

[9] Sample replicates were divided into thirds and new sample was released at 5:00, 5:30, and 6:00 either Eastern Standard Time or Central Standard Time depending on the location of the number. Replicates were released early on to allow at least one callback on each number (the calling period ended at 9:00PM).

successfully contacted in the multi-day IVR poll even though 21,277 of those numbers were called three times.[10]

[TABLE 1 ABOUT HERE]

The calculated CASRO response rate for our human poll survey is 16.7%. Given the difficulties in identifying the disposition of numbers in the IVR poll noted above, it is impossible to directly compute a response rate for the IVR polls. Because the same sampling frame was used to generate the numbers being called in each poll, we can approximate the response rate by comparing the ratio of the number of completes to the number of numbers that were called. For example, in our human interview poll, 573 completes were obtained after calling 21,258 unique numbers – a ratio of 2.7%. Our multi-day IVR survey had 1,248 completes after calling 49,986 unique numbers (2.5%), and our one-day IVR poll had 1204 completes from 49,991 unique numbers (2.4%). All else is equal – which is an admittedly strong assumption that likely overstates the performance of IVR polls with their static and relatively impersonal introduction – comparing these ratios would suggest that the response rate of the IVR polls is, at best, 92.6% that of a human landline poll.

These calculations and comparisons depend critically on whether callbacks are made. If we only count completes produced by the initial call, the ratio of completes to numbers for the multi-day IVR poll would be 1.67%, and the ratio for the human poll would be 2%. These ratios suggest that IVR response rates for a poll without a callback is, at best, 83.5% of a poll conducted by human interviewers without callbacks.

Although callbacks improve the response rate of IVR polls, they have diminishing returns if the IVR technology cannot screen out ineligible and invalid numbers. On the first evening of

[10] An implication of this is that IVR surveys with call backs may not be very efficient or cost-effective because most IVR companies charge per call.

our multi-day IVR poll, for example, 16,666 numbers were called and 388 completes were obtained – a ratio of completes to numbers of 2.3%. On the second evening, 13,139 of these numbers were recalled, but just 0.7% of these calls produced a complete. The completion rate fell even further to only 0.3% on the third evening. Overall, our multi-day IVR poll made 86,000 callbacks and yielded only 296 additional completes. In contrast, 837 completes were obtained in the first 50,000 calls.

Another important difference between a human interview poll and an IVR poll is the significant drop in the willingness of a respondent to complete an IVR survey. Only 54 respondents (8.6%) failed to complete the twenty-minute long survey with human interviewers, but nearly 40% of those who started our three-minute long IVR polls failed to complete the interview.

[FIGURE 1 ABOUT HERE]

Figure 1 depicts the precipitous drop-off in IVR poll cooperation across our three minute long survey by graphing the percentage of respondents answering each question given that they listened to the poll's introduction and they verified they were at least 18 years old. Across both IVR polls, over 3,600 respondents listened to the pre-recorded 26 second introduction and then answered that they were at least 18 years old. However only 82% of these individuals answered the next question on presidential approval, and participation continues to decline throughout the brief interview. In fact, only 61% of the respondents who start the survey complete the survey, and the largest drop-off occurs after respondents were asked to enter their zip code or year of birth.[11]

2. Who Responds?

[11] The 82% of individuals who only answered the first two questions is comparable to the findings reported by Competitive Edge Research & Communication (Neinstedt 2011).

Having demonstrated that human and IVR interviews produce different response rates even when using the same sampling frame, we now compare who responds to each mode and how the respondents to each of our polls compares to the target population according to the Census or American Community Survey.[12] Table 2 describes the respondents of our three surveys relative to the current Census estimates for the state of Tennessee. Characteristics in which the sample proportion differs significantly from the adult population are denoted with asterisks, and instances in which the IVR respondents differ from the human-interviewer respondents are so noted.

[TABLE 2 ABOUT HERE]

Table 2 reveals that all three sets of respondents differ significantly from the population of adult Tennesseans. Reflecting known issues in landline samples, those who complete the surveys are more likely to be female and older than the state as a whole. Nearly 50% are above the age of 65 regardless of the survey mode employed and the percentage of respondents who are under the age of 45 is extremely small – especially on the IVR surveys. 47% of the adult state population is under the age of 45, but less than 12% of respondents to IVR polls were similarly aged (our human poll was slightly better – 17% of the interviews were given by those under the age of 45).

As Table 2 makes clear, our landline polls differ significantly from statewide estimates in nearly every demographic category. Moreover, the last two columns of Table 2 reveal that the demographic biases of landline-only surveys are exacerbated even more in the two IVR polls we

[12] One difference between the two survey modes that is worth noting is that while we randomize the selection of household members when conducting a human interview by asking to speak to the household member with the most recent birthday, such randomization is not easily done when conducting an IVR poll. As a result, IVR poll respondents are household members who answer the phone. If the people who answer the phone in a household are more engaged than those who do not, holding fixed demographic characteristics, then this could be consequential if the unobservable characteristics are uncorrelated with the observable characteristics.

conduct. IVR polls are even less likely attract young people than human conducted landline polls, and they are also less likely to contain respondents who possess a high-school diploma or less. These are large groups in the population, so their relative paucity can pose problems for estimating public opinion – especially if the opinion of interest divides along these lines.

3. Differences of Opinion?

Ideally, poll respondents would perfectly represent the target population and weighting would be unnecessary. This is obviously not the case in our experiment. Because the characteristics of the poll respondents differ from the demographics of the target populations, weighting adjustments are required if we are interested in using the sample to try to learn about opinion in the state. However, it is difficult to assess the quality of a poll when only weighted top-line results are reported - as is commonly the case for publically reported IVR polls - precisely because weighting adjustments can have such important effects on the estimates of public opinion.

To compare how estimates of public opinion vary between our three surveys in light of the evident demographic imbalances, we adopt a common weighting strategy for each. In particular, we use an iterative weighting procedure (implemented via the package rake in R) to produce weights for individual respondents based on minimizing the sum of the differences in actual and target values based on gender, age, education, and region of the state (DeBell and Krosnick 2010).[13] We compute sample weights that are both "trimmed" so that the maximal weight is capped at 5, and "untrimmed." In both cases, the weights for each raking variable sum to 1.

[13] Specifically, in addition to gender we weight on a three-category age variable – 18-44 ; 45-64 ; 65+ – a three-category education variable – high school or less ; some college ; college and post-graduate – and three category region variable – East ; Middle; West. The groupings were chosen so that at least 5% of the sample were in each cell using IVR respondents.

Before proceeding to a description of political attitudes, it is useful to consider how the decision to trim or not influences the individual weights for IVR poll respondents.[14] Figure 2 plots the relationship between the trimmed and non-trimmed weights for the one-day IVR poll (the relationship in the multi-day IVR poll is similar), and it shows how trimming truncates roughly 2.6% of the individual weights. Given the disparities evident in Table 2, the weights that are most affected are those for males under the age of 45. In contrast, the decision to trim or not has no consequence for the landline human poll.

[FIGURE 2 ABOUT HERE]

Given these weights, we can now consider how the weighted estimates of public opinion for the three polls differ. This is important because the weights we construct arguably correct for the differences we document in Table 1 (and even characteristics that are highly correlated with the measures), but there may also be unobservable characteristics that affect the likelihood of a respondent completing an IVR poll that affect public opinion.[15]

Two analyses are of interest: 1) how do weights affect the estimates for each survey and, 2) how do estimates of public opinion differ after weighting the surveys using an identical procedure? The latter comparison is important because it reveals whether IVR and human polls produce different estimates holding constant the sample of telephone numbers called, the wording and order of questions that are asked, and the nature of the weighting procedures employed.

---

[14] As further evidence of the sample differences, there are significant differences in the estimated design effects for the three polls. The general design effect is 1.92 for the trimmed and untrimmed weights for our human interview poll, but it is 3.21 for the untrimmed one-day IVR, 2.70 for the trimmed one-day IVR, 4.09 for the untrimmed multi-day IVR, and 3.11 for the trimmed multi-day IVR.

[15] Speculatively, perhaps those respondents who are most willing to take a survey via IVR are those who are exceptionally interested or engaged in the subject and who do not therefore need a human on the other end to encourage them to complete the survey. Our ability to pursue this inquiry is limited by the constraints imposed by a three-minute survey, but we can take the first step of determining whether there are differences in public opinion between the survey modes even after employing individual weights.

Figure 3 presents the results graphically by illustrating the proportion answering affirmatively for every question common to the three polls and the 95% confidence interval for each estimate and reveals several findings.[16] First, even after weighting, there are statistically distinguishable and politically consequential differences in the estimated proportions across the three surveys. Overall, of the 8 questions, there are 5 statistically distinguishable differences between the human poll results and the one-day IVR poll and 4 statistically distinguishable differences between the human poll and the multi-day IVR poll. In fact, there are distinguishable differences between the two IVR polls in 4 of the 8 questions that are asked.

Second, the decision to use weights is clearly consequential. Focusing on the multi-day IVR poll, for example, reveals differences of 5% in the approval of the state legislature and 6% in the approval of Governor Haslam depending on whether weights are used. These differences can obviously be consequential in the context of close election polls, but given the unrepresentativeness of who responds, we focus on the differences in weighted estimates because those estimates attempt to eliminate demographic differences.[17]

Third, there are systematic differences between the weighted results. Figure 3 reveals that one of the largest differences is in the proportion of respondents claiming to be an "independent." Whereas nearly 37% of respondents tell a human interviewer that they are independent, only 26% and 27% respond accordingly on our one-day and multi-day IVR polls, respectively.[18] IVR respondents are far more likely than respondents to a human poll to self-

[16] Tables A1 and A2 in the appendix report the point estimates and whether there are statistically significant differences between the two.

[17] There are also significant differences in the estimated design effects between the polls that could prove important. While our analyses in Table 3 did not adjust the standard errors to account for the design effects due to weighting reported in footnote 12, so doing would decrease our ability to say anything definitive about public opinion using IVR polls because our uncertainty about the point estimate would increase substantially.

[18] Similarly, Lohuizen and Samohyl (2011) find that Republican primary voters who respond to IVR polls are more likely to report being "Very conservative" instead of "Moderate/Liberal," and Blumenthal and Franklin (2007) find higher levels of undecided voters amongst IVR samples.

identify with a major party (66% in an IVR poll versus 51% in a human poll), and there are more self-identified Democrats in particular (35% in the multi-day IVR poll; 30% in the one-day IVR poll; and 23% in the human poll). Fewer respondents also self-identify with the Tea Party when talking to a human (16% report being a member of the Tea Party to a human interviewer, but the proportion rises to 21% and 22% for multi-day and one-day IVR polls respectively).[19]

[INSERT FIGURE 3 ABOUT HERE]

There are also important differences in respondents' opinions about political actors. Obama's overall approval rating, for example, ranges between 30% (human poll) and 35% (multi-day IVR), but striking differences emerge if we examine presidential approval by age. Among respondents who were at least 40 years old, Obama's average approval does not differ much between the three polls – ranging from 30% (human poll) to 35% (one-day IVR). Among respondents under the age of forty, however, there are striking differences -- Obama's average approval rating among these respondents ranges from 25% (one-day IVR) to 40% (multi-day IVR). Moreover, even with over 1,000 respondents to our multi-day IVR poll, the 95% confidence interval for Obama's average approval ranges from 26% to 55% for those under the age of forty. Because of this imprecision, we cannot say anything conclusive about how Tennesseans under the age of 40 think of Obama -- we cannot even determine whether or not a majority approves of his performance.

Finally, there are also differences between modes in the proportion of respondents who approve of the performance of state-level political figures. Overall, 59% of respondents to the human poll approve of Republican state legislature's performance, but only 53% and 47%

[19] As the survey instrument in Appendix A clarifies, we ask identification with the Tea Party separately from partisan self-identification to allow for membership in the Tea Party to span across political parties. This can be especially important in a state like Tennessee where the meaning of the parties has evolved so dramatically over time.

approve in the one-day and multi-day IVR polls respectively. Governor Haslam's (R) approval in the multi-day IVR poll is 58% versus 64% and 65% in the human and one-day IVR polls. Putting aside the difference in Haslam's approval rating, respondents living in Tennessee are generally less likely to support liberal political figures such as President Obama and more likely to approve of conservative political leaders when being interviewed by a human relative to those that are interviewed by an IVR poll.

Given the nature of the questions being asked, we can only speculate as to why these differences might exist, but one reason might be that members of partisan out-groups are less willing to admit their partisanship in a so-called "red state" to a human interviewer because of social desirability bias or a desire to conform to the perceived community (Tourangeau, Steiger, and Wilson 2001; Kreuter, Presser and Tourangeau 2008). If so, the desirability to conform to the norms of the perceived community when asked by a fellow human being may result in more satisfaction being expressed relative to a survey that is self-administered and the pressure to self-identify with the "in-group" may be weaker. Such a reaction would constitute clear evidence of a mode effect (Couper 2011).

Alternatively, perhaps members of the political out-groups are more motivated than most to respond to an IVR study to express their displeasure than those who are relatively content with the status quo.[20] That is, perhaps those who are most motivated to express their displeasure with the status quo are willing to take a self-administered pre-recorded poll, but there are some individuals who would decline to participate in a self-administered poll that will participate if the survey is administered by a human being. Perhaps suggestive of the latter, in a review of the performance of pre-election polls following the 2010 midterm elections in which Republicans

[20] To be clear, while we weight based on observable characteristics, the claim is that there are unobservable characteristics such as motivation that cannot be controlled for and which differ across the group of respondents.

retook the House of Representatives from the Democrat majority, Silver (2010) documents that IVR polls were more Republican-leaning than other polls. Similar to the pattern that we find, perhaps those that are most displeased with the status quo are the most motivated to complete an IVR poll. If so, this may speculatively – and also problematically – suggest that the desire to participate in an IVR poll depends on the underlying political context and that the differences we observe are due to a selection effect in who responds.

While we cannot adjudicate between whether the differences we highlight are due to selection or mode effects (e.g. Rivers and Vavreck 2012), further study is critically important because how we interpret public opinion depends on which is more consequential. If personal human-interviews are more likely to produce results biased by desirability effects because of mode effects than the opinions of political minorities and outgroups would be understated in such surveys because of their desire to express opinions that are shared by others. If however, those same groups are the most likely to respond to automated, self-administered surveys to express their displeasure and selection effects result, then the results of IVR surveys would likely overstate the opinions of the outgroup because of their increased participation rate in the survey. Further complicating the issue is the possibility that these mode and selection effects are not mutually exclusive and that both effects may be present.

4. Conclusion and Implications

Political polling has increasingly begun to dominate the media's coverage of politics. Given the amount of attention being paid to the results of polls and the increase in the number and types of polls being conducted, it is perhaps more important than ever to understand the differences between different polling technologies. Past performance alone is not enough to justify the validity of a poll (e.g. Burns 2010; Zengerle 2012); just as we would not use the fact

that the incumbent president's party won the presidential election every time the Washington Redskins won their final NFL home game between 1936 and 2004 to justify its use as a forecasting method, we cannot use the past performance of a poll alone to validate its quality without also exploring its strengths and weaknesses to assess its possible future shortcomings.

To better understand the costs and benefits of using IVR polls to assess public opinion on important political matters, we conduct an experiment where we survey the same population using the same survey instrument using different interviewing modes. This design allows us to identify the similarities and differences in response rates, sample demographics, and opinion estimates between landline samples conducted using human interviews and IVR technology.

We find that IVR polls have a lower response rate than a human-conducted poll and that IVR response rates are affected by a significant drop-off in participation even on a brief three-minute poll. Moreover, the demographic biases that are known to affect landline-only surveys are exacerbated in our two IVR polls. For example, those who complete our IVR polls are even more likely to be old and female than a landline survey conducted by human polls.

In addition to differences in who responds, we also find important differences in public opinion depending on the mode of interview even after weighting. Respondents to our IVR survey in the largely conservative state of Tennessee are more likely to self-identify with a political party – including the Tea Party – and they are also more approving of liberal political figures than the estimates of the human poll would suggest.

We cannot determine why these discrepancies occur, but the differences we document suggest that care should be taken when interpreting and comparing IVR poll results to human poll results. IVR polls can certainly be done faster and more cheaply than polls using human interviewers, but we find that the public opinion we estimate using our IVR polls of varying field

periods are only sometimes similar to the results of a near-identical human poll conducted under nearly identical conditions. Since these differences are larger than we would expect due to chance alone, we cannot disentangle whether the differences are due to mode effects or selection effects. Interpreting discrepant results produced by the two modes of interviewing requires caution until we better understand the trade-offs associated with each.

The measurement and reporting of public opinion has taken an increasingly large role in recent political coverage. It is therefore critical that the estimates of public opinion that are being reported accurately reflect public opinion. There are many methods of measuring public opinion, but only when we understand the potential biases of a polling methodology well-enough so that we can properly account for the possible errors should we be confident that it can capture what the public thinks about an issue or a candidate. Given the rapid increase in the number of polls relying on Interactive Voice Recording technology we provide an important starting point for this critical line of inquiry. Our results suggest caution is warranted, but much work remains.

References

AAPOR. 1997. *AAPOR Finds Frank Luntz in Violation of Ethics Code*. [press release] April 23, 1997.

An Evaluation of the Methodology of the 2008 Pre-Election Primary Polls. American Association for Public Opinion Research, March 30, 2009. http://aapor.org/uploads/AAPOR_Rept_FINAL-Rev-4-13-09.pdf.

Anand, Anika. "Polling Company Fires Back at NBC, Chuck Todd." Salon.com, 2010. http://www.salon.com/2010/06/30/ppp_on_nbc/.

Belli, Robert F., Michael W. Traugott, and Matthew N. Beckmann. 2001. "What Leads to Voting Overreports? Contrasts of Overreporters to Validated Votes and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics* 17:479–98.

Blumberg, Stephen, and Julian V. Luke. Wireless Substitution: Early Release of Estimates From the National Heath Interview Survey, January-June 2013. Center for Disease Control and Prevention, December 2013.

Blumenthal, Mark, and Charles Franklin. "Methods and Horse Races: Phone, IVR, and Internet Polls in the 2006 Elections." presented at the AAPOR, http://www.pollster.com/blogs/Franklin_Blumenthal_2007_AAPORPresentation.pdf, 2007.

Blumenthal, Mark M. "Toward an Open-Source Methodology What We Can Learn from the Blogosphere." *Public Opinion Quarterly* 69, no. 5 (January 1, 2005): 655–669. doi:10.1093/poq/nfi059.

Bougher, Lori and Markus Prior. "Political Polarization of the American public continues to rise. Or does it?" Washington Post (October 3,2013).

http://www.washingtonpost.com/blogs/monkey-cage/wp/2013/10/03/political-polarization-of-the-american-public-continues-to-rise-or-does-it/

Burns, Alexander. “Florida’s Other Winner: Robo-Polling.” Politico, August 25, 2010. http://www.politico.com/news/stories/0810/41455.html.

Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. "The effects of mode and format on answers to scalar questions in telephone and web surveys." *Advances in telephone survey methodology* (2008): 250-275.

Clinton, Joshua D., and Steven Rogers. “Robo-Polls: Taking Cues from Traditional Sources?” PS: Political Science & Politics 46, no. 02 (2013): 333–337. doi:10.1017/S1049096513000012.

Cohen, Jon. “Covering Automated Surveys in 2012.” The Washington Post - Blogs, February 16, 2012. http://www.washingtonpost.com/blogs/behind-the-numbers/post/covering-automated-surveys-in-2012/2012/02/01/gIQANvmJIR_blog.html.

Cohn, Nate. “PPP’s Results Don’t Excuse Its Bad Methodology.” The New Republic, September 19, 2013. http://www.newrepublic.com/article/114769/ppp-methodology-results-arent-defense.

———. “There’s Something Wrong With America’s Premier Liberal Pollster.” The New Republic, September 12, 2013. http://www.newrepublic.com/article/114682/ppp-polling-methodology-opaque-flawed.

Couper, Mick P. “The Future of Modes of Data Collection,” *Public Opinion Quarterly* 75.5 (2011)” 889-908.

Debell, M, and J.A. Kosnick. Computing Weights for American National Election Study Survey Data, 2010. ftp://ftp.electionstudies.org/ftp/nes/bibliography/documents/nes012427.pdf.

DiCamillo, Mark. “The Rise of Robopolling in California in 2010 and Its Implications.” The Polling Report, November 15, 2010. http://www.pollingreport.com/md1011.htm.

Elliott, Justin. 2010. “Research 2000 Prez Lashes Out At Kos In Detailed Response To Fraud Suit.” *Talking Points Memo*. July 2. http://talkingpointsmemo.com/muckraker/research-2000-prez-lashes-out-at-kos-in-detailed-response-to-fraud-suit.

Gallup, George, and Saul Forbes Rae. *The Pulse of Democracy: The Public-Opinion Poll and How It Works*. Westport, Conn.: Greenwood Press, 1973.

Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. "Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias." *Public Opinion Quarterly* 67.1 (2003): 79-125.

Keeter, Scott. “Pew Research Increases Share of Interviews Conducted by Cellphone.” Pew Research Center, January 15, 2014. http://www.pewresearch.org/fact-tank/2014/01/15/pew-research-increases-share-of-interviews-conducted-by-cellphone/.

Kennedy, Courtney, and Stephen E. Everett. "Use of cognitive shortcuts in landline and cell phone surveys." *Public opinion quarterly* 75.2 (2011): 336-348.

Kos. “Daily Kos: Fear and Loathing of ‘Robo-Polls’ at the Washington Post.” Accessed January 27, 2014. http://www.dailykos.com/story/2011/08/09/1004682/-Fear-and-loathing-of-robo-polls-at-the-Washington-Post.

Line, Laura H., Michael F. Schober, Frederick G. Conrad, and Heidi Reichert. 2013. “Why Do Surey Respondents Disclose More when Computers Ask the Questions?” *Public Opinion Quarterly* 77.4 (2013): 888-935.

Lohuizen, Jan van, and Robert Wayne Samohyl. "Method Effects and Robo-Polls." *Survey Practice* 4, no. 1 (June 14, 2013). http://surveypractice.org/index.php/SurveyPractice/article/view/92.

Malhotra, Neil, and Jon A. Krosnick. "The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples." *Political Analysis* 15.3 (2007): 286-323.

Moulitsas, Markos. "'Robo-polls' don't lie." Text. TheHill, July 19, 2011. http://thehill.com/opinion/columnists/markos-moulitas/172363-robo-polls-dont-lie.

Neinstedt, John. "84% Hung up on Robo-Poll: Another Test Raises Serious Questions." The Edge: Competive Edge Blog, May 16, 2011. http://cerc.net/2011/05/16/84-hung-up-on-robo-poll-another-test-raises-serious-questions/.

Parry, Hugh J., and Helen M. Crossley. "Validity of responses to survey questions." *Public Opinion Quarterly* 14.1 (1950): 61-80.

Preisendörfer, Peter and Felix Wolter. "Who is Telling the Truth?: A Validation Study on Determinants of Response Behavior in Surveys," *Public Opinion Quarterly*. First published online February 19, 2014 doi:10.1093/poq/nft079.

Presser, Stanley, and Linda Stinson. 1998. "Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance." *American Sociological Review* 63:137–45.

Rivers, Doug, and Lynn Vavreck. "Estimating Mode Effects without Bias: A Randomized Experiment to Compare Mode Differences between Face-to-Face Interviews and Web Surveys," The American Association for Public Opinion Research (AAPOR) 67th Annual Conference, 2012.

Rosentiel, Tom. “Cell Phones and Election Polls: An Update.” Pew Research Center, October 13, 2010. http://www.pewresearch.org/2010/10/13/cell-phones-and-election-polls-an-update/.

Silver, Nate. “‘Robopolls’ Significantly More Favorable to Republicans Than Traditional Surveys.” FiveThirtyEight. Accessed January 27, 2014. http://fivethirtyeight.blogs.nytimes.com/2010/10/28/robopolls-significantly-more-favorable-to-republicans-than-traditional-surveys/.

———. “Which Polls Fared Best (and Worst) in the 2012 Presidential Race.” FiveThirtyEight. Accessed January 27, 2014. http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/.

Stern, Michael J., Ipek Bilgen, and Don A. Dillman. "The State of Survey Methodology: Challenges, Dilemmas, and New Frontiers in the Era of the Tailored Design." *Field Methods* (2014): 1525822X13519561.

Tourangeau, Roger, Darby Miller Steiger, and David Wilson. "Self-administered questions by telephone." *Public Opinion Quarterly* 66.2 (2001): 265-278.

Ye, Cong, Jenna Fulton, and Roger Tourangeau. "More positive or more extreme? A meta-analysis of mode differences in response choice." *Public opinion quarterly* 75.2 (2011): 349-365.

Yeager, David S., et al. "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples." *Public Opinion Quarterly* 75.4 (2011): 709-747.

Zengerle, Jason. “The Polls Ultimately Ended Up Making Sense — But Next Time, Who Knows?” *Daily Intelligencer*. Accessed January 27, 2014. http://nymag.com/daily/intelligencer/2012/11/polls-in-the-end-ended-up-making-sense.html.

Tables

	MultiDay IVR	OneDay IVR	Rate (OneDay)	Human	Rate (Human)
Numbers Called	49986	49991		21258	
Non-Residential	Unknown	Unknown		756	
Computer Fax	Unknown	Unknown		689	
Other Not Working	2164	2150		15734	
Working Numbers	47822	47841	95.7%	4079	19.2%
Contacted Numbers	10524	10525	22.0%	2203	54.0%
Callback				140	
Refusal				1319	
Cooperating Numbers	1934	1934	18.4%	744	33.8%
Eligible Numbers	1818	1808	93.4%	627	84.3%
Completes	1132	1078	59.6&	573	91.4%
Partial Completes	686	730	40.4%	54	8.6%

Table 1: Sample Final Disposition for Long Run IVR, Short Run IVR, and Human Poll: Dispositions are reported by the vendors we use (Precision Polling for the IVR polls and Princeton Survey Research Associates International for the Human Poll).

Category	Census Estimate	Human Poll	One Day IVR	Multiday IVR	One Day sig. different from Human	Multiday sig. different from Human
N		573	1078	1132		
Female	0.52	0.59*	0.62*	0.65*	No	Yes
18-24	0.12	0.02*	0.01*	0.00*	No	Yes
25-34	0.17	0.05*	0.03*	0.02*	Yes	Yes
35-44	0.18	0.09*	0.08*	0.07*	No	No
45-64	0.35	0.35	0.38*	0.39*	No	No
65+	0.17	0.49*	0.51*	0.51*	No	No
High School or less	0.51	0.42*	0.36*	0.32*	Yes	Yes
College Degree	0.15	0.20*	0.15	0.20*	Yes	No
Grad. Degree	0.08	0.13*	0.15*	0.15*	No	No
Urban	0.38	0.25*	0.29*	0.31*	No	Yes
Rural	0.23	0.28*	0.29*	0.27*	No	No
Suburban	0.38	0.47*	0.42*	0.42*	No	No

Table 2: Composition of Respondents: * Reflects Poll Proportion different from Population Proportion ($p \leq .05$, two-tailed).

Figures

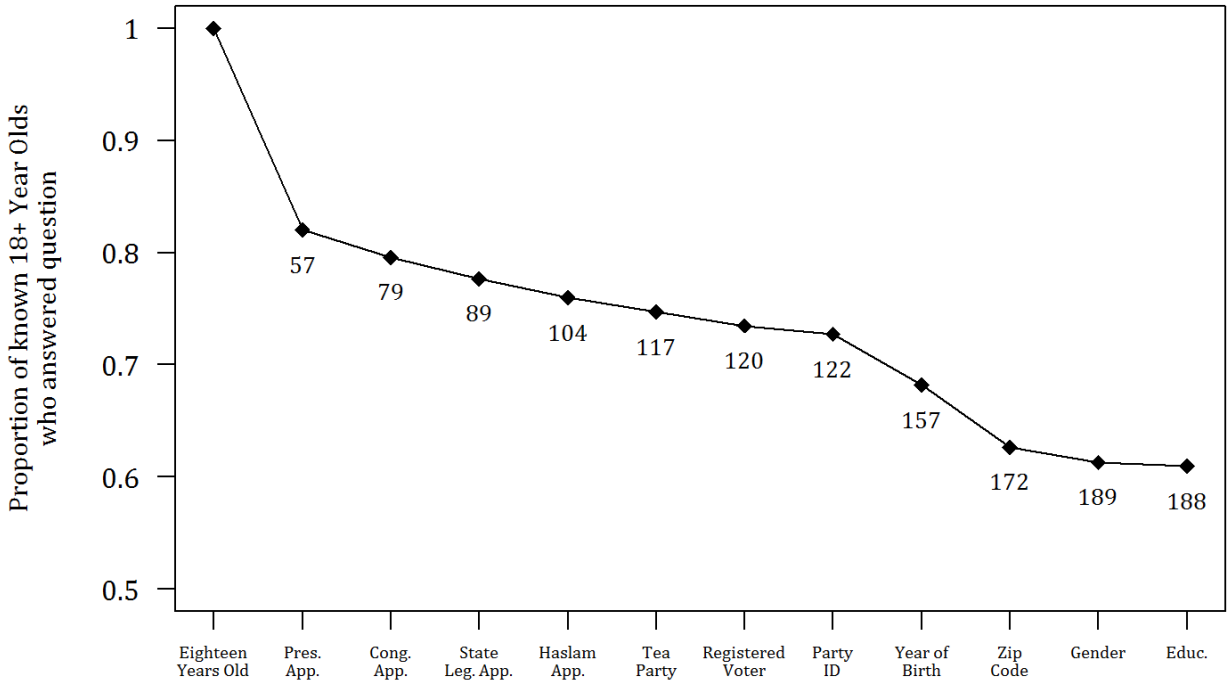

Figure 1: Proportion of Respondents Answering Each Question in IVR Polls: Proportion of respondents who answer each question conditional on listening to the IVR introduction and verifying they were 18 years of age or older. Numbers below plotted proportions reflect the average length of a call in seconds for respondents who hung up immediately before the indicated question.

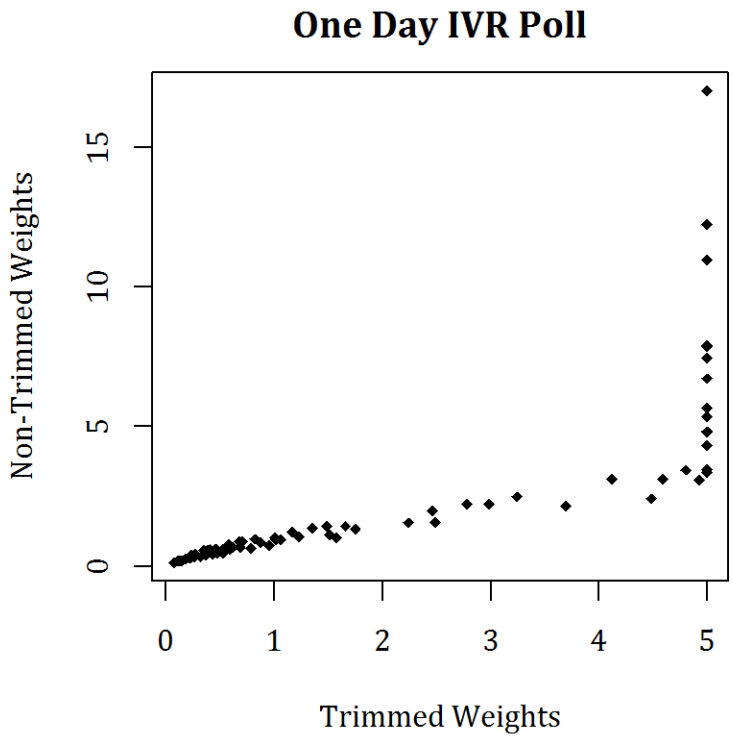

Figure 2: Comparing Trimmed and Non-Trimmed Weights for One-Day IVR Poll: Reported weights are based on gender, a three category age variable, a three category education variable, and a three category region variable using the iterative raking procedure implemented in rake in R.

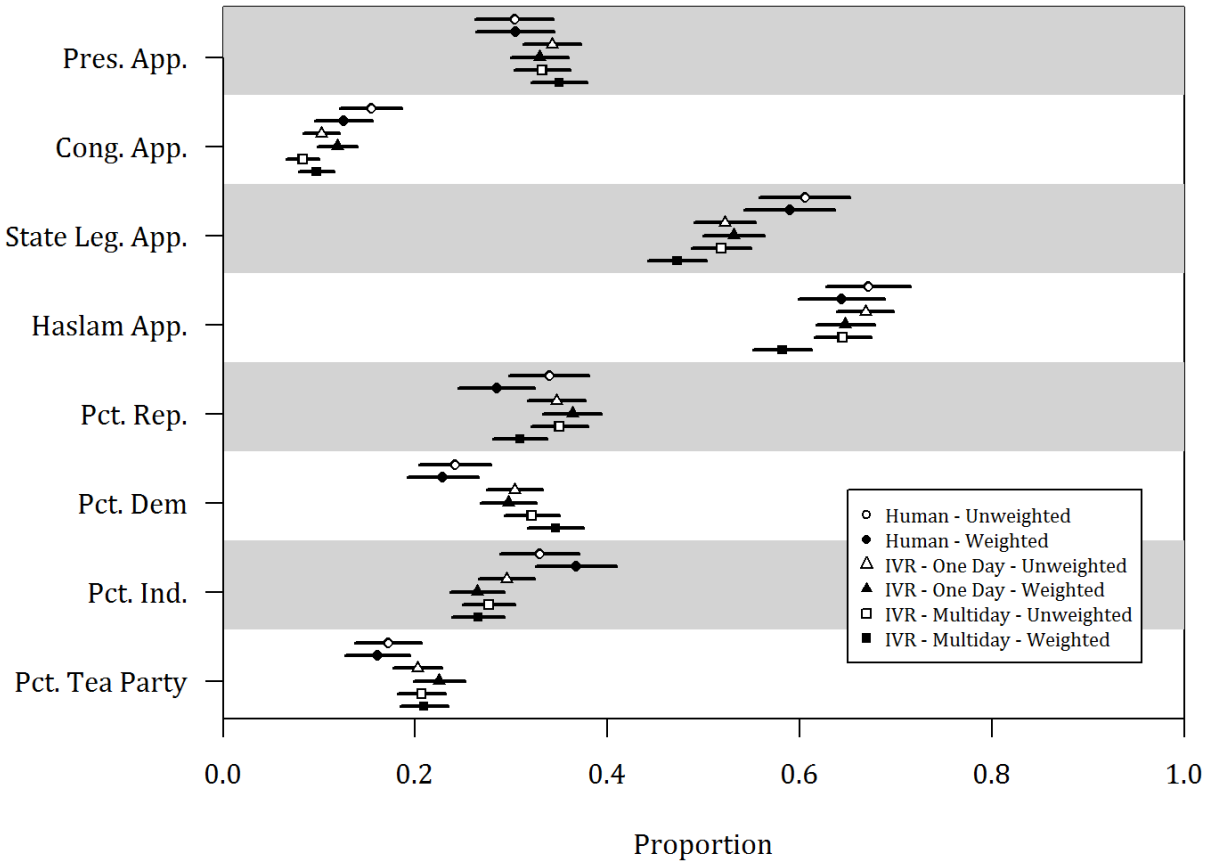

Figure 3: Proportion Answering "Yes" to Each Common Question: Line segments denote 95% confidence intervals. Weighted (trimmed) and unweighted estimates are reported for each. The impact of design effects are omitted.

Appendix: IVR Poll Text -- November 2013

Hello, my name is _____ and I'm calling for Vanderbilt University. We're conducting a survey about some important issues today, and would like to include your household. Your phone number has been randomly selected to complete the survey. We are not selling anything.

Q0 Are you at least 18 years of age and currently living in the state of Tennessee?
1. Yes
2. No

If Q0=2.

"Thank you for your time, but we are only interested in the opinions of Tennesseans at this time."

If Q0=1:

Q1 I'd like your views on some government organizations and elected officials. As I read each, please tell me if you approve or disapprove of how each is handling their job. Do you approve or disapprove of:

a. The job Barack Obama is doing as President
b. The job the U.S. Congress is doing
c. The job the Tennessee State Legislature is doing
d. The job Bill Haslam is doing as Governor

CATEGORIES
1 Approve
2 Disapprove

Q2 Do you think of yourself as part of the Tea Party movement?

1 Yes
2 No

REGIST These days, many people are so busy they can't find time to register to vote, or move around so often they don't get a chance to re-register. Are you NOW registered to vote in your precinct or election district or haven't you been able to register so far?

1 Yes, registered
2 No, not registered

PARTY Generally speaking, do you usually think of yourself as a:

1 Democrat
2 Republican
3 Independent (OR)
4 Something Else

“Finally, we have some questions that are used for statistical purposes only.”

AGE What year were you born _____ ?

ZIP What is your zipcode _____?

GENDER Are you:
1. Male
2. Female

AGE What was the last grade of school you completed?
1. No High School Diploma
2. High School Graduate
3. Some College or Associate Degree
4. College Graduate
5. Postgraduate study

CLOSE “THANK YOU again for sharing your thoughts and opinions!

Table A1 – Weighted Means and Differences

	Human Mean	One Day IVR Mean	Multiday IVR Mean	Sig. diff btwn. IVR Polls
Pres. Approval	0.30	0.33	0.35	No
Cong. Approval	0.13	0.12	0.10	No
State Leg Approval	0.59	0.53*	0.47*	Yes
Haslam Approval	0.64	0.65	0.58*	Yes
Republican Pct.	0.28	0.36*	0.31	Yes
Democrat Pct.	0.23	0.30*	0.35*	Yes
Independent Pct.	0.37	0.26*	0.27*	No
Tea Party Pct.	0.16	0.22*	0.21*	No

* Reflects IVR Proportion different from Human Poll Proportion ($p \leq .05$, two-tailed). Sample Self-Reported Registered Voters.

Table A2 – Unweighted Means and Differences

	Human Mean	One Day IVR Mean	Multiday IVR Mean	Sig. diff btwn. IVR Polls
Pres. Approval	0.30	0.34	0.33	No
Cong. Approval	0.15	0.10*	0.08*	No
State Leg Approval	0.60	0.52*	0.52*	No
Haslam Approval	0.67	0.67	0.64	No
Republican Pct.	0.34	0.35	0.35	No
Democrat Pct.	0.24	0.30*	0.32*	No
Independent Pct.	0.33	0.29	0.28*	No
Tea Party Pct.	0.17	0.20	0.21	No

* Reflects IVR Proportion different from Human Poll Proportion ($p \leq .05$, two-tailed). Sample Self-Reported Registered Voters.