

**Enough Already with Random Digit Dialing:  
A Proposal to Use Registration-Based Sampling to Improve  
Pre-Election Polling**

Donald P. Green  
Yale University

Alan S. Gerber  
Yale University

May 5, 2002

Abstract: Pre-election polls invariably use random digit dialing (RDD) to draw representative samples of likely voters. We make the case for an alternative sampling methodology, clustered random sampling from voter registration lists. Because these lists often furnish useful background information that strongly predicts whether a person will vote, registration-based sampling (RBS) may provide more accurate election forecasts. This paper proposes to test the relative accuracy of RDD and RBS in predicting election outcomes.

Paper prepared for Presentation to the Gallup Conference on Improving the Accuracy of Polling, May 2-4, 2002, Washington, DC. The authors are grateful to the Institution for Social and Policy Studies at Yale University for research support. We would also like to thank David Nickerson, who gathered the registration and voter turnout data used here.

## **Enough Already with Random Digit Dialing: A Proposal to Use Registration-Based Sampling to Improve Pre-Election Polling**

Pre-election polling is a highly standardized activity. The overwhelming majority of pre-election surveys rely on telephone interviewing, and the overwhelming majority of phone surveys use random digit dialing, whereby calls are placed to randomly generated phone numbers within pre-selected phone exchanges. Because phone calls are placed to people whose identities and addresses are unknown, the survey analyst must rely on the respondent to furnish all of the key pieces of information needed to interpret the survey results. The respondent must help the interviewer select a randomly chosen member of the household for questioning, who in turn must furnish information about whether they are currently registered to vote and whether they intend to vote in the upcoming election.

Because the aim of pre-election polling is to describe the opinions of the voting public, many respondents are excused during the interview process. Unregistered voters receive a polite thank you, and their interviews are terminated early on. Respondents who make it past this screen are cross-examined about their proclivity to vote in the upcoming election. Those who express uncertainty about whether they will cast a ballot or who report having skipped elections in the past are discreetly removed from the sample afterwards. Those who remain constitute the sample of “likely voters” whose opinions (perhaps after a bit of massaging using weights for the respondents’ party affiliation or demographic profile) become the poll’s official results.

Whether this procedure provides a reliable means for forecasting election outcomes depends in part on whether the interviewer accurately discerns which respondents will vote. In many, if not most elections, voters and nonvoters have different candidate preferences; mistaking a nonvoter for a voter in these instances will bias the survey results. Pollsters are, of course, well aware of this problem. When pre-election polls mispredict election outcomes, pollsters and poll-watchers invariably lay the blame at the feet of voter turnout. This ritual is in some ways convenient for pollsters because it means that they need not adjust their survey procedures in order to account for it. The vicissitudes of voter turnout are regarded as a sort of uncontrollable force of nature for which pre-election forecasters are not responsible.

Polling organizations could do a far better job of anticipating voter turnout by changing one facet of their standard operating procedure. We propose replacing random digit dialing (RDD) with registration-based sampling (RBS), which is neither difficult nor costly to implement. Indeed, the advent of database management firms, the digitization of records by local registrars, and pending legislation designed to standardize this record-keeping on a statewide basis makes RBS an increasingly attractive alternative to RDD. Although we can make a strong case in the abstract for the advantages of using RBS, whether registration-based sampling leads to improved election forecasts remains an empirical question, and the purpose of this paper is to make the case for why this alternative sampling methodology warrants further investigation.

## The Allure of RDD

During the 1970s, the polling industry began moving away from face-to-face interviewing and embraced phone surveys. The advantages of phone surveys were many, such as lower unit costs, easier supervision, and shorter lead times. At that time, longstanding skepticism about the representativeness of telephone surveys began to subside. The overwhelming majority of American residences had phone service. There remained, however, the problem of drawing a random sample of the population. RDD sampling is premised on the idea that one cannot obtain a representative sample of households using listed phone numbers. Some people, after all, have unlisted phone numbers, and excluding them from the sample leads to bias.

This argument persuaded the polling world to embrace RDD and put up with various problems specific to it. Although the phone digits are drawn randomly, the people who answer the phone are not. RDD requires the pollster to wade through large numbers of nonworking or nonresidential numbers in search of residences. Once a residence is identified, an accomplishment that requires some kind of contact with a person or an informative answering machine, the pollster must draw a random sample within the household. This task involves either some form of enumeration (e.g., "How many registered voters are living at this address?") or some near-random selection criteria (e.g., "May I speak to the adult who will be having the next birthday?"). Executing this type of within-household random sampling may run into practical difficulties stemming from the suspicion or puzzlement raised by these odd questions.

Even when within-household sampling proceeds without incident, there remains the task of re-weighting the sample to reflect the sampling probabilities. In simple random sampling, each observation has the same probability of being selected. When RDD sampling is clustered by telephone exchanges, these weights are determined by the putative number of residential numbers within each exchange. Another problem is that some residences have more than one number, which means that the pollster must inquire about the number of phone lines in the household that could be used in a survey conversation (e.g., phone lines not dedicated to a computer modem). The final survey data are then re-weighted according to the probability of selection into the sample.

Were the survey simply attempting to draw a representative sample of the general population, these complications might be worth putting up with in order to reap the main benefit of RDD, namely, overcoming the problem of unlisted residential numbers. RDD seems to do a reasonable job of sampling the adult population. But when the aim is to draw a representative sample of those who will vote on Election Day, the merits of RDD become more ambiguous.

## Sampling the Population of Actual Voters

Ideally, a pre-election poll is a random draw from the population of people who will actually cast ballots on Election Day. To be sure, minds may change between the time the poll is conducted and when ballots are cast, but error associated with opinion

change is conceptually distinct from the error associated with sampling from a population other than that of actual voters. Even before Election Day, we have quite a bit of information about who will vote. The population of actual voters is a subset of the population of registered voters. With the exception of a few states that have same-day voter registration or lack registration requirements, the population of voters is drawn from the population of those who have registered at least a few weeks before the upcoming election. Incidentally, in states with traditional registration systems, only a tiny fraction of the voting electorate registers during the months leading up to an election. For example, in Raleigh, North Carolina, 7% of the 264,669 voters who cast ballots in 2000 had registered after October 1, 2000. In Seattle, the corresponding figure is 4% of 786,286.

Knowing which voters in an RDD survey are registered well in advance of an election gives us a leg up on the problem of describing the population of those who will actually cast ballots. In fact, many pre-election polls conducted a few months prior to an election are content to describe the population of "registered" as opposed to "likely" voters. However, when one uses RDD, one typically relies on the respondent to indicate whether he or she is registered. One could conceivably perform a reverse phone match linking the RDD number to a registration list, but the phone numbers on registration lists tend to be outdated, many registrars do not require phone numbers, and the respondent must furnish address information in order to find the right registration list. How accurate are respondents' self-reports about their registration status? No one knows.

Figuring out who is registered is really just the first step toward knowing who will actually vote. Here, too, we must rely on the representations made by the respondent. The typical pre-election survey peppers respondents with questions about the frequency with which they voted in the past and their likelihood of doing so in the upcoming election. Based on this series of questions, the survey analyst makes some kind of determination about which respondents constitute "likely voters." (Alternatively, the analyst could assign each respondent a probability of voting, but this does not seem to be the industry practice.) Survey results are tallied for this pool of likely voters, and that's that.

The problems with this approach are twofold. First, respondents may misstate their voting intentions. Second, the survey analyst is generally at a loss to translate even a sincere report of past or intended voting behavior into a probability of voting in the upcoming election. Naïve classification of these responses into rigid "likely voter" and "non-likely voter" categories only compounds the problem, since it discards meaningful variation within these groups. Taken together, these problems mean that the "likely voter" designation is really just guesswork.

In principle, the accuracy of this designation could be improved by rigorous study of how expressed vote intentions predict actual voter turnout. Such validated vote studies

are rare, even among academic surveys such as the American National Election Studies.<sup>1</sup> Validated vote studies conducted by commercial firms doing pre-election phone surveys are rarer still. To its credit, the Gallup Poll attempted a voter validation study after the 1992 election based on an RDD sample, but it ran into serious operational problems. In order to match respondents to voter registration data, it first needed to get respondents' names and addresses. This proved difficult, and only a minority of cases were actually matched.

This uncertainty leads to our first proposal: a vote intention and vote validation study. The design of this study is very simple. Draw a sample from a list of registered voters, conduct a phone interview, and examine the relationship between vote intention measures and actual voting behavior. By working from a registration list, one makes trivial the costs of validating voter turnout. Presumably, this basic study could be embellished by replicating it in various electoral settings (e.g., general vs. primary elections) and elections for different types of offices (e.g., national, state, municipal). Results would provide for the first time some means of calibrating survey responses and actual voting rates.

### Registration Based Surveys

Voter validation is an important advantage of working from registration lists, but it is not the main advantage. Registration data contains a wealth of information that may be used to forecast voter turnout. Although the quality of information available from public records varies across jurisdiction, the typical registration file contains date of birth, date of registration, and party registration. In addition, one may obtain data on past voter turnout. Sometimes these voter histories are furnished by registrars, and sometimes by private vendors. The reliability of these data doubtless varies, but the point remains that one can produce a powerful statistical prediction of future voting behavior simply by reference to public records -- before one even speaks with a respondent. (Indeed, working from registration list helps answer some nagging questions about the biases associated with survey nonresponse, since we would know quite a bit about the respondents who could not be interviewed.)

The powerful relationship between public information and voter turnout is illustrated by voting patterns in Raleigh and Seattle. Suppose we were interested in forecasting voter turnout in the 2000 general election. Our voter history files contain two very useful pieces of information: when the voter registered and whether he or she voted in the 1998 general election and the 2000 primary. Table 1 presents the cross-tabulation of voter turnout in 2000 by whether the individual was registered and voted in the preceding 1998 and 2000 elections. Table 1 shows a powerful relationship between a voter's prior vote history and their subsequent voter turnout. For example, in Seattle we

---

<sup>1</sup> One might be inclined to turn to American National Election Study surveys, but these have not performed vote validation in recent years, and older surveys tended to rely on in-person interviewing, which may have different misreporting rates than phone surveys.

see that among the 411,526 voters who were registered in 1998 but skipped both the 1998 and 2000 primary elections, turnout in the 2000 election was a dismal 31.0%. By contrast, those who voted in both these previous elections (N=326,302) voted at an astonishing rate of 97.4%.

Intuition suggests that any survey of the voting electorate should place special emphasis on the opinions of citizens with a high *ex ante* probability of voting. Consider the limiting case, for example, in which members of the population come in two types, those who are certain to vote and those who are certain to abstain. There would be no point in interviewing any members of the latter group. By this logic, constructing an optimal sample – that is, a sample that gives the smallest prediction errors when forecasting the actual election outcome – should take notice of prior voting history because it powerfully predicts whether someone will vote in the upcoming election.

The current RDD sampling procedures are a long way from optimal. Pollsters simply discard the interviews they conduct with those they deem to be unlikely voters. That does not sound like a cost efficient procedure, but RDD does not give them much choice, because the pollster has no idea which respondents are likely to vote before interviewing them. RBS, on the other hand, gives pollsters a means of differentiating between likely voters and nonvoters *before* money is spent interviewing them.

The actual procedure of drawing an optimal sample is complicated a bit by the fact that the candidate preferences of voters with ardent partisan preferences are easier to predict than those with moderate views. Knowing that a voter is a regular participant in Republican primaries leaves relatively little uncertainty about how he or she will vote. For this reason, the pollster need not interview as many regular primary voters as voters in other equally sized centrist groups in the population. The rule of thumb, which is formalized mathematically in the appendix, is this: allocate more interviews to large subgroups of voters, unless you know ahead of time how they are likely to vote. For example, in Seattle, primary voters who also voted in 1998 comprised 41.5% of the 2000 electorate. This group requires a bit of attention due to its enormous size. On the other hand, a pollster will quickly learn after a few interviews that the GOP stalwarts favor Bush and that the Democratic regulars favor Gore. One could invest additional interviews in this group, but those interviews are better spent on other large groups about which there is more uncertainty. The next largest group, those who voted in 1998 but skipped the 2000 primary, comprised 26.6% of the 2000 electorate. This group commands extra polling resources.

Of course, in advance of an election one cannot know how large each group's share of the electorate will be. It is therefore handy to know how voting history has predicted subsequent turnout in recent elections. Presumably, before attempting to construct an optimal survey design for the 2002 election, a pollster might consult the relationship between turnout in the 1996 and 1998 elections. It should be stressed, however, that getting the rates just right is not crucial. Knowing simply that one group is much more likely to go to the polls than another helps a pollster allocate resources more

efficiently than would be the case with simple random sampling. Even if the eventual sample is not optimal, it is still a big step in the right direction.

To see how RDD and RBS might compare in terms of sampling error, consider a hypothetical election in which the Republican candidate wins 52.5% of the vote against a Democratic opponent. Suppose that the adult population consists of two groups, each of which comprises half the population. Group one is 55% Republican and turns out to vote at a rate of 90%. Group two is 45% Republican and votes at a rate of just 30%. Suppose this population were to be sampled using two types of sampling frames, RDD and RBS. Each survey consists of 1000 completed interviews. The reported RDD results, however, are restricted to the sample of self-professed likely voters. Suppose that 90% of those who will actually vote claim an intention to vote, whereas 35% of those who will not vote claim an intention to vote. Computer simulations of this scenario demonstrate that the likely voter/RDD sample has an average prediction error of 1.8 percentage-points, as compared to 1.2 for the RBS sample. (Note that the RBS poll made no use of the stated vote intention for purposes of this illustration.) RDD incorrectly predicts a Democratic victory 32% of the time, as compared to 8% for RBS. In sum, the RBS sample of 1000 completed interviews is as accurate as a corresponding RDD sample of 2200 completed interviews.

Nothing prevents the RBS analyst from combining registration data with stated vote intention in order to come up with an even more accurate forecast. The empirical question is whether survey responses predict subsequent behavior over and above the statistical information available in registration data. No one knows the answer to this question, and this research topic should be a central part of the voter validation study described earlier. Perhaps voters do convey useful information about their subsequent behavior when they claim that they intend to vote. Or perhaps these claims are made to mollify the interviewer and add nothing to the statistical prediction of the respondents' behavior. The extent to which survey responses improve our predictions of what people actually do is an empirical question. We strongly suspect, however, that public documents will prove to be the more reliable source of information.

Although we have used the term “registration” to refer to voter registration lists, the idea behind RBS extends to any sort of list-based sampling frame. For example, in polities without voter registration, postal lists could be used as the basis for sampling. The statistical advantages of RBS, however, stem from the ancillary information about the prospective respondent that one gathers before interviews are conducted. Voter registration lists at a minimum tell us the voter's age, date of registration, and whether other voters are registered at the same address – all important predictors of voter turnout. Moreover, registration lists include address, which enables the pollster to find out a great deal about the political and economic climate in which the respondent lives and, by extension, how the potential respondent is likely to behave on Election Day.

RBS in Practice

Registration-based sampling starts with a registration list. These lists are, with very few exceptions, available from one or two sources. The first source is the registrar of a local jurisdiction, such as a city or county. Although we cannot claim to have done an exhaustive study of the extent to which these local registrars work with digitized lists, we have experience with local registrars. Our voter mobilization experiments have made us pen pals with registrars across the country -- in small rural communities, suburbs, and large cities. We have yet to encounter a registration list that was not in machine-readable form. Typically, all that is involved in obtaining this list is a small fee, although sometimes one must also promise not to use the list for telemarketing. In some cases, these lists are officially off-limits to anyone who is not connected with a political party. Parties seem to be quite porous, however, and these forbidden lists are commercially available from any number of database vendors. These vendors charge more for these data than local registrars, but the prices are still quite modest. The advantage of purchasing these data from vendors is that they have often taken the trouble to append to the dataset information about the previous elections in which a person has voted; registration lists obtained directly from registrars may not contain this information, although it can often be obtained from them separately in machine-readable form. Moreover, many vendors have access to the most recently updated databases of residential phone numbers.

Matching addresses and phone numbers is the point at which slippage occurs in the RBS sampling process. The quality of the match has in our experience varied widely by region, and of course nothing prevents the pollster from falling back on RDD in cases where the phone match is particularly poor. It should be stressed, however, that the forecasting accuracy of RBS in comparison to RDD is an empirical question, for as we note below, RBS may have certain advantages in terms of reducing nonresponse. If RDD phone surveys are enlisting the cooperation of less than half of the households they contact, it is by no means clear ex ante which source of bias is more problematic. The only way to gauge the relative accuracy of RDD and RBS is to do both and see what happens. Another fruitful line of research would be to examine whether the accuracy of election forecasts across jurisdictions varies with the proportion of unlisted numbers.

Which jurisdictions' registration lists one must obtain is a matter of clustered probability sampling. Here the standard logic of clustered sampling applies. Create primary sampling units, and choose randomly among them. Then break the primary units into secondary sampling units, and so forth. Eventually, one gets down to small clusters, such as a Census block group. If, for example, one were to select 50 clusters with 50 names apiece, one would have a working sample of 2,500 names, which would be ample to support a phone survey of 1,000 completed interviews even assuming a mediocre response rate. While gathering registration data from these clusters is a nuisance (although a nuisance that database vendors would happily embrace, for a fee), creating a bit of sampling infrastructure has its benefits. One can sample in the future from clusters located in these towns, merely updating the existing database with new registrants and new voter histories. One could even contemplate panel studies or analyses that made use of contextual information provided by other respondents living in the same local cluster.

Whereas RDD uses telephone exchange clusters, RBS uses geographic clusters. Geography can in certain circumstances have important advantages. After redistricting, phone surveys may have a difficult time locating respondents who live within a newly crafted House district. The respondents themselves may not know which district they live in, and exchanges may spill over into other districts. Address information, on the other hand, can be fed into GIS software containing the boundary files of a new district.

Once the RBS sample of respondents has been selected, the mechanics of the survey are quite straightforward. Since the name of the respondent is known to the interviewer, the caller may ask for the respondent by name. No further enumeration of the household is necessary. Whether calling the respondent by name improves rapport between interviewer and respondent remains an empirical question. If not, one could always fall back on the RDD protocol, which never mentions the respondent's name. It may be that respondents in RBS surveys need extra assurances about confidentiality, although this does not seem to be a special concern in face-to-face surveys, where similar issues arise.

In terms of gaining the cooperation of respondents, one special advantage of RBS is that it enables survey firms to send out a letter ahead of time explaining the purpose of the survey, inviting participation, and perhaps offering some incentive. A further advantage is that it gives the survey firm an alternative to telephone interviewing in the event that the respondent refuses an interview. Presumably, one could offer the respondent who initially refuses the alternative of taking an internet-based survey. No such fallback position is available in RDD surveys because the only means of contacting respondents is via telephone. In an era when phone surveys are confronting serious nonresponse problems, these kinds of considerations weigh heavily in favor of RBS.

A final advantage of RBS is that it enables pollsters to contact more efficiently hard-to-find populations. In most states, where absentee voters constitute a small but important fraction of the electorate, it would be prohibitively expensive to conduct a survey of absentee voters. Similarly, it would be costly to attempt a survey of young voters or voters who have recently switched parties. RBS seems particularly valuable for journalists or political marketers, who seek to study small segments of the electorate.

## Conclusion

Random digit dialing shares much in common with the QWERTY keyboard. Both inventions grappled with an important practical limitation at the time they were conceived. The QWERTY keyboard resolved the problem of typewriter keys sticking together when the typist went too quickly. RDD overcame the problems of unlisted residential phone numbers and spotty machine-readable lists of eligible voters. Both QWERTY and RDD have become nearly universal, but neither has adapted to other technological changes. Computer keyboards can accommodate much faster typing speeds, and digitized registration lists abound. There is no end in sight for the QWERTY keyboard because so many people are used to working with a certain layout of keys. The investment in RDD is different. Polling firms make extensive use of RDD, but they also

conduct surveys of specific populations, such as consumers or business executives. There is more room for change, and some polling firms might welcome the opportunity to get out from under the inconvenience of phoning numbers at random.

We are not calling for an end to RDD, even within the domain of political polling. Our argument is confined to pre-election forecasts and, even then, our thesis is simply that this new idea deserves to be subjected to rigorous testing. Let the data tell us which sampling methodology leads to the most accurate election forecasts. The results may come down squarely in favor of one across a range of different elections, or it may be that RBS is found to be especially effective in, say, low turnout elections. Time will tell whether RBS is indeed a better mousetrap.

Table 1

## Voter Turnout in 2000 by Voter Turnout in Two Previous Election, Raleigh and Seattle

<b>Raleigh</b>	Vote 2000	<b>Not Registered in 1998</b>			<b>Abstained in 1998</b>		<b>Voted in 1998</b>	
		Not Registered for 2000 Primary	Abstained 2000 Primary	Voted 2000 Primary	Abstained 2000 Primary	Voted 2000 Primary	Abstained 2000 Primary	Voted 2000 Primary
Voted in 2000 General Election	NO	12,053 (34.7%)	44,159 (67.3%)	236 (5.6 %)	60,631 (49.8%)	363 (6.0%)	12,137 (11.7%)	1,363 (2.3%)
	YES	22,622 (65.3 %)	21,483 (32.7 %)	3,867 (94.4%)	61,038 (50.2%)	5,691 (94%)	91,738 (88.3%)	59,139 (97.7%)
<b>Seattle</b>	Vote 2000	<b>Not Registered in 1998</b>			<b>Abstained in 1998</b>		<b>Voted in 1998</b>	
Voted in 2000 General Election	NO	33,990 (37.4%)	25,779 (46.4%)	1,088 (8.8%)	283,860 (69.0%)	3,284 (8.9%)	35,609 (14.6%)	8,401 (2.6%)
	YES	56,976 (62.6%)	29,791 (53.6%)	11,311 (91.2%)	127,666 (31.0%)	33,764 (91.1%)	208,877 (85.4%)	317,901 (97.4%)