*Science*, Nov 28, 1986 v234 p1094(5)

# Assessing the Accuracy of Polls and Surveys

*Philip E. Converse; Michael W. Traugott.*

IN 1984 RONALD REAGAN CAPTURED 59% OF THE NATION'S popular votes, and of the electoral votes save those in challenger Walter Mondale's home state. Published pre-election polls generally picked President Reagan as the likely winner. Yet even late in the campaign, quite discrepant estimates of the victory margin were appearing. At the extremes, a Gordon Black survey conducted for USA Today gave Reagan a lead over Mondale by 60% to 35% with 5% undecided, while a Roper Poll for the Public Broadcasting System showed Reagan ahead 52.5% to 42.5%, Also with 5% undecided (1). Earlier polls, even when simultaneous, had diverged still more widely.

Discrepancies as glaring as these are not common for reputable sample surveys of the same populations at the same time, but they do occur. And while polls reported in the national media tend to state error margins, usually plus or minus three percentage points, a reader diligent enough to compare competing polls, especially in presidential election seasons, is likely to conclude that error margins must somehow exceed this three percentage point value by an appreciable amount.

It is important to be realistic about the precision of sample surveys because their results have been given increasing weight in our national life in recent decades. In the 1980 presidential campaign, for example, the League of Women Voters set a threshold of 15% popular support "in the polls" for aspirants to qualify as participants in its televised presidential debates. Less formal, but full as important, is the recognition among campaign strategists that rising poll support typically means a greater flow of dollars into the candidate's campaign coffers, just as erosion of support predicts a collapse of contributions.

Nor, of course, is the power of the polls limited to the campaign season. No major national issue--be it the defense budget, our role in Nicaragua, or tax reform--is debated any longer without reference to numerous public opinion polls reported in the mass media. Clearly these poll results do not in any direct way determine governmental policy, but it is also obvious that they are watched with care by a significant segment of the nation's decision-makers.

Indeed, it is the perceived power of the polls that has in recent years prompted the development of "advocacy polling," which is the use of polls by interest groups to create an aura of strong popular support for their favored positions. Typically this is achieved by the employment of unrepresentative samples, leading questions, or selective reporting of results. Such "findings" merely add to the sense of cacophony in national polls and surveys.

When published polls disagree, several casual reactions are possible. One is to blame sampling error: while large discrepancies are not expected to occur with any frequency, they will in fact occur from time to time. Another reaction is that public opinion is known to be volatile, and since it is rare that competing polls have been carried out on exactly the same days, perhaps discrepancies

are due to real changes in preferences. Still another is to decide that public opinion is too amorphous for measurement, and hence its results are best disregarded.

There is probably some grain of truth in all of these reactions. However, each taken alone is far too extreme; and all of these reactions together fail to represent very adequately the full range of reasons why polls can diverge. The sheer proliferation of sample surveys covering the same topics in recent years has drawn a good deal of attention to the variability of estimation in such work, including scrutiny by two National Research Council panels under the auspices of the Committee on National Statistics, one devoted to problems of incomplete data (2), the other to difficulties in the measurement of subjective states(3).

Concerns of this kind have generated notions of "total survey error," which attempt to encompass not only sampling error, but additional sources as well (4). These additional sources include at least two major forms of missing data, that due to limits on sample coverage and to nonresponse; and several forms of measurement error traceable to the interviewer, the respondent, or the questionnaire. Even a brief examination of these sources of error may help explain why survey results can diverge.

**Sampling Error**

The standard warning label giving error margins for published poll results typically refers only to the most obvious source of variability, the error arising because the population at issue has not been fully enumerated, but merely sampled. The margins are based on the sampling error of a proportion (the square root of [p(1 - p)/N]), known to approximate a normal distribution when N is in the ranges customary for such samples.

The usual 3% warning represents the rounded evaluation of the confidence interval for a population proportion (p) of 0.50, at a probability value of 0.95, when the sample is of the relatively conventional size of 1500 cases. With progressively smaller samples the error margins may be widened to four or five percentage points, for the chief governor precision for such estimates, comparable to the resolving power of telescopes, is indeed sample size. A three percentage point warning for a sample of 1500 cases may in one sense be seen as conservative, since the unrounded solution is only [plus-or-minus]2.53 percentage points, and even this value takes the error at its maximum (where p = 0.5): once past very even divisions of the population, the calculated error margin drops below [plus-or-minus]2.50%.

In practice, however, the three percentage point warning is not conservative even for "pure sampling error" because its calculation presumes simple random sampling. But such an elementary sample "design" is prohibitively expensive to implement in surveys of human populations, whether one interviews by telephone or by personal visits to dwelling units. More feasible but complex sample designs employ one or another form of geographic clustering, such as by census tract for personal interviewing or by telephone exchange.

Such clustering, while cost-effective, inflates the appropriate error calculation in the degree that the attribute being estimated is relatively homogeneous within clusters, as family income tends to be by residential neighborhoods (5). Another awkward consequence of clustering is the fact that sampling error varies from one attribute to another in the same sample. Thus a general warning such as the conventional three percentage points is at best a crude average and can conceal wide variation in expected error across attributes. Most attributes,

including opinions on political issues, are typically not so maldistributed across convenient clusters as to increase error margins by 30 or 40%, and often suffer as little as 5% increases over simple random sample error. On the other hand, for some common "geographic" attributes, such as urban-rural residence and its close correlates, the inflation produced by clustering can easily triple or quadruple the expected error.

Finally, a minimal requirement for any probability sample designs that every member of some bounded population has a known nonzero chance of being selected in the survey. This is the requirement that permits an estimate of sampling error. A significant fraction of published polls--even those marked "scientific"--have been executed without the costs involved in any formal probability selection model. For these surveys, sampling error is not only unknown but unknowable, and any error margin cited has only metaphoric status at best.

With all of this in mind, the conventional three percentage point warning is rarely a conservative one, even for "mere sampling error taken alone. But sampling error is thought of as the variability in estimates to be expected from replications of the sample design with freshly drawn samples, while holding other conditions essentially constant. It is only one form of variable error in surveys, and ignores as well the possibility of biases which are fixed error. Total survey error thus expands the concept of error from those derived from sampling error alone (6).

**Comparing Products of Different Survey Houses**

Experience with observed discrepancies between surveys underscores the fact that sampling error alone tends to be an inadequate explanation. At the same time, this can be somewhat harder to demonstrate than appears. The problem is that a discrepancy well in excess of the proclaimed three percentage point margin discovered between results published for two surveys is bound to catch the eye and create suspicion. Yet the same theory which says discrepancies this large are unlikely to occur fluctuation also implies that upon occasion, such as one time in a hundred or a thousand, rather larger discrepancies will arise on sampling grounds alone. Therefore it is not too compelling to seize upon some discovered discrepancy of six percentage points as "beyond plausible sampling error" when in fact it has come to attention through an implicit scan over hundreds of less" noteworthy" comparisons.

The most meaningful comparisons, therefore, must be based on a fuller distribution of discrepancies, rather than on stray examples. Since different survey agencies rarely pose the same questions at the same times, such distributional contrasts are not easy to assemble. Smith (7) has, however, located 33 instances with the same item being asked of a national sample in the same general period by some pair of four survey agencies, two of them university-based (the Survey Research Center at the University of Michigan and the National Opinion Research Center at the University of Chicago) and two commercial (Roper and Gallup's American Institute of Public Opinion). He found statistically significant differences ($p < 0.05$) for estimates of the same proportion in 10 of the 33 comparisons.

While these results suggest "house effects" which cannot be dismissed as mere sampling error, they are less conclusive than might appear. For one thing, in order to amass as many as 33 comparisons it was necessary to include surveys that were in the field as much as five months apart. Furthermore, there was a systematic positive association between the length of the timing displacement and the size of the absolute discrepancy in results. Thus, for example, for the eight comparison

synchronized to within 2 months or less, none showed a significant difference. Five of the ten significant differences were clustered in one battery of linked comparisons displaced by an intermediate number of months; and the other five occurred over the 14 instances in which 4 months or more elapsed between the measurements. Even with these massed comparisons, true change over time can scarcely be discarded as an explanation for discrepancies that appear to exceed sampling error.

Despite a paucity of truly tight comparisons, there remains significant evidence that discrepancies do arise between matched surveys which exceed sampling expectations in size and frequency (3, 8). Kiewiet and Rivers (9), for example, have done a "meta-analysis" of results of presidential "trial heats" published by nearly a dozen commercial polling organizations during the 1984 campaign. This body of data contains a greater density of close temporal matches than are usually found elsewhere, and is of particular interest because it is these discrepancies in presidential preferences that are most likely to be noticed by the casual reader and to breed skepticism about polls more generally. The analysts show that "house" differences unquestionably exceed plausible sampling error and, furthermore, display signs of persisting bias, with the same houses repeatedly overestimating or underestimating the Reagan margin, with timing of measurement controlled.

It must be noted that voting comparisons are bound to exaggerate house effects, in comparison with the general case, because different houses use quite different means of running such trial heats. Furthermore, they rarely publish "raw" data as they come from the field, insisting instead upon inserting various weights (the nature of which is usually guarded as a proprietary secret of the house) thought to compensate for such things as known sample flaws and differential voter turnout. Nonetheless, sufficient traces of such house bias have been found in other contexts that few seasoned researchers would place a subjective 95% confidence interval as narrow as a mere three percentage points around assorted survey estimates.

Substantial variation in measurement outcomes from one research agency to another is scarcely novel even in the annals of natural science (10, 11). But the urgent practical question becomes one of proper allowances to be made for such nonsampling errors. Does prudence require that we subjectively quadruple the confidence intervals suggested for pure sampling variability? Or will a modest expansion by one-quarter or less suffice?

It is unrealistic to expect some reliable general-purpose answer. Certainly the lay reader who cannot detect advocacy polls in a stream of publishes results would be well advised to imagine much wider error margins that those who can discard some results as intrinsically questionable. But as we shall see, other problems arise even for sophisticated readers, due to less visible forms of quality variation in the complex measurement process called survey research.

**Variability in Sample Composition**

It is obvious that we may expect discrepant results from surveys that sample from different frames or lists. Some political surveys are based on samples of all adults, for example, while others are based on all registered voters, or even "likely" voters. Even where the conceived sample frame is identical, however, various practical problems arise in the course of implementation of the design which produce shortfalls from the ideal. The two most obvious of these are (i) under coverage because the frame population is not a full listing of the target population and

(ii) nonresponse from eligible sample members (2).

Most national surveys claim in shorthand to cover the adult population in the United States. However, virtually all surveys in the private sector miss a small margin of the adult population. Few surveys include Alaska or Hawaii, or institutionalized members of the population in hospitals, barracks, dormitories, and jails. Interviewing by telephone is less expensive than by personal visit, and the vast majority of surveys published in the media are now conducted by telephone. But despite what is called telephone "saturation," some 7 to 8% of the household population remains inaccessible by residential phone. Sample designs for personal interviewing usually revolve around the dwelling unit, and thus have a fuller coverage, but still systematically miss some of the poor, the rural, and the transient.

These biases in coverage provide a good example of the "fixed error" portion of total survey error, since presumably they are constant over replications of the same design. Such errors will not be apparent in house comparisons for which the same general procedures are used. However, different modes of interviewing, such as telephone versus personal, do have different coverage and hence can produce discrepancies in results. This fact has led to suggestions for a mixing of modes when high precision is required (12), although major costs may be entailed.

Much more important for potential error is the problem of nonresponse. There are two broad types, according to whether persons designated by the sample design to be interviewed fail to answer phones or doorbells or are successfully contacted but refuse to provide an interview. Nonresponse appears more serious than under coverage because the shortfall typically caused is much larger. For personal interviewing, nonresponse has more than doubled in private sector work during the past four decades, with levels even for careful field operations now approaching 30%. Most of this increase has come in the refusal component; and since it has been sharpest for the most urbanized areas, it has been presumed to reflect a growing unwillingness to open doors to strangers (13, 14). Initially it was hoped that interviewing by telephone might represent some solution to this problem. However, while it is not easy to match response rate calculations between the two modes, it now appears that nonresponse rates tend to run higher for telephone interviewing than for personal visitation.

There is an obvious association between effort invested and improved response rates. In the normal case in which the sample design does not leave latitude to interview any household member, not much more than one-third of the interviews can be completed with a single call, either for telephoning or visiting, despite the optimizing of time of approach. Each successive round of callbacks adds to the response rate but also to study costs. Efforts to convert initial refusals even by offers of payment for cooperation or mailed explanations of the importance of the study have parallel cost and benefit implications. In practice, the effort invested, and hence the response rate, varies remarkably by agency according to philosophy and within agency by the funding level for any specific study. For most of the "timely" polls appearing in the media, short deadlines and presumed volatility of response require that measurement be completed in a very short interval, such as 1 to 3 days. This severely limits the possibility of callbacks or the efficacy of those that are made. In the face of these pressures, the hastier polls freely substitute other accessible people for designated respondents who cannot be found quickly, or they completely abandon

probability designs that designated specific respondents.

The overall bias contributed by nonresponse depends logically on levels of nonresponse and the degree to which nonrespondents differ from those interviewed in relevant respects. The first is a single parameter that can often be estimated; the second is a whole family of parameters, one for each variable of concern, most of which remain unknown. The exceptions include the few attributes like sex and approximate age that can be observed in the course of a direct refusal or those that can be deduced aggregatively from shortfalls in sample characteristics relative to census distributions for the population. Once past the reliable finding that response rates are lower in urbanized areas, studies of nonresponse seem to generate quite mixed results, although more often than not the old and the less educated seem to show higher probabilities of nonresponse (15, 16).

Studies that employ numerous callbacks permit analysis of variability as a function of apparent inaccessibility. This is more satisfying in the sense that any variable in the study can be examined for bias, although conclusions concerning nonresponse bias hinge on an arguable assumption that those never interviewed are distinctive in a manner similar to those less easy to interview. It is not uncommon to discover that respondents who require several callbacks differ on variables central to the survey (17) in comparison with those interviewed at the first call. In one of our political surveys during the 1984 campaign, for example, Democratic partisans were more accessible at early calls than Republican ones. A trial heat gave Reagan a mere three percentage point margin over Mondale among those interviewed at one call; those answering a callback helped increase the lead to six percentage points; for the final sample, after up to 30 callbacks for the most difficult to locate respondents, the lead had advanced to 13 percentage points (18). Thus callback policy differences have a potential for creating discrepancies in survey results, whatever the status of "hardcore" nonrespondents who are essentially impossible to find or to cajole into cooperation.

It is not uncommon that published survey results are "weighted" in some manner in an effort to minimize the effects of any known bias due to nonresponse. Thus if a sample proves to have a lower ratio of men to women than census data show, male cases may be given a higher proportional weight than female ones so that the sex ratio in the sample is brought to the known parameter. Adjustments of this kind are rarely communicated in media publications and can represent another hidden source of variability in results. On the other hand, the attributes for which such adjustments can be made are few, and it is the common experience that most variables of interest are too poorly correlated with demographic differences to vary palpably even when such weights are applied (19).

More generally, while sample composition must account for some fraction of nonsampling error, it is likely that more severe problems arise in the course of the measurement process itself.

**Measurement Variability: Interviewers**

Among common variations that exist between survey agencies are those which have to do with the amount of interviewer training given, both as to general procedures and in preparation for implementing the instruments specific to given studies. Interviewing staffs also differ in rates of turnover and hence accumulated experience. Training and seasoning influence many parts of the survey process. For example, a well-trained staff of interviewers, winnowed over time through

performance reviews, will generate higher response rates for given levels of effort.

What is seen as good interviewer practice can also vary by agency. When systematic comparisons can be made on opinion items posed by different agencies at essentially the same time, the "house effect" most commonly found lies in different proportions of "don't know" responses (8). It seems clear that some agencies encourage their interviewers to push their respondents to some substantive response, rather than accept a "don't know," whereas others point out that this may produce artificial responses. In making cross-agency comparisons, it is usually wise to remove such non-content responses from the distribution, and the fact that this is one of the most predictable forms of "house effect" is an indirect comment on the general robustness of substantive responses.

How much extraneous variance in item responses can be traced to the assignment of one interviewer rather than another to a given respondent? It is established that an interviewer who betrays personal feelings that some responses are more sensible than others will in fact garner more of the "desirable" responses. On the other hand, one of the main goals of interviewer training is to foster a nondirective style of interviewing that keeps such effects to a minimum. For the most part, such training appears to succeed. Concern lingers, however, with respect to subtle expectations set up will-nilly by the interviewer's sex, age, race, or other manifest characteristics. the effect most reliably demonstrated is that both black and white respondents report positions on race-related issues that are less supportive of blacks when talking to white interviewers than when talking to black ones (3, 20, 21). Such discrepancies by racial pairing disappear, however, on matters of opinion not racially sensitive, and efforts to find effects associated with other kinds of interviewer-respondent pairing have usually shown little.

All told, although the inevitable intrusion of human interviewers on the measurement process can only add to nonsampling variability, experience suggests that, given proper training, the addition is usually trivial, although systematic differences ("house effects" may not be. Nonetheless, training is a frequent target for cost "saving."

Measurement Variability: The Questionnaire

For carefully run sample survey operations, it is likely that the greatest potential for significant unforeseen nonsampling variability resides in the way in which the questionnaire is constructed. The two main sources of variability that are well documented involve the choice of question wording and the interview context in which items are lodged (22). Wording may vary according to selection of words in the item or more generically by the form in which a question is cast. One of the basic watersheds of question form is that between "open" questions, which invite respondents to shape answers in their own terms, and "closed" questions, which require selection among a set of fixed answers. For purposes of speed of administration and standardization, most surveys now published in the media restrict themselves to closed questions. Even within the domain of closed questions, however, a myriad of form variations exists, in such matters as the number and style of alternatives offered, whether defaults like "don't know" or "no opinion" are presented explicitly, and the like.

Opinions that seem to be the "same" can be measured quite differently, and often, although not always, with quite different results. Thus, for example, when in early 1979 the SALT II treaty approached its Senate vote, many polling agencies sampled national opinion on the subject. The naive

reader paging from one to another set of results would find them to discordant as to produce doubt that true opinions on the subject were even measurable: one cluster of polls showed 2 to 1 majorities favoring such arms limitation, and another cluster showed majorities roughly 2 to 1 against. A more careful examination showed that two broad types of wording were being used: one asked about the desirability of strategic arms limitation in principle; the other tied the issue to the specifics of SALT II. In short, roughly a third of the public wanted some SALT treaty, but not the one at hand. In effect, the respondents were being more reliable and discriminating than casual readers or, perhaps, the investigators themselves.

In the political realm, some crucial matters such as levels of support for particular candidates might seem to be immune from such variability of measurement. There is, in fact, a great deal. It is obvious that the support for any given candidate will vary according to the opponent presumed, although pairings in such "trial heats" are usually identified as part of new reports. More subtly, results vary in some degree upon such things as whether a party affiliation or incumbency status is made explicit in the question, or even the order in which a series of trial heats is presented.

Experienced practitioners rarely place weight on "discrepancies" between questions that are worded differently. They may be less sensitized, however, to the possibility that responses to questions identically worded can vary upon occasion because of the context in which they are asked. In the case of candidate support, for example, there is reason to believe that an incumbent candidate will draw warmer responses if assessed after a series of questions that highlight dimensions of performance where he has been successful than after items that evoke, however inadvertently, more negative outcomes. Or again, respondents were less likely to report that they followed politics closely after a series of grueling information questions concerning their local political representatives than in less embarrassing contexts (23).

The contexts for most questions are essentially neutral, and when some particular context turns out to have produced a bias in responses, it is usually not hard to see why. However, context bias can easily arise inadvertently; and, of course, most readers of published polls cannot evaluate the context possibility because the content of the full questionnaire is rarely mentioned or explicitly provided.

Variations in question wording and context seem to produce more nonsampling variability in results than discrepancies in sample composition or interviewer effects. For the most part, of course, questionnaire decisions are innocent ones. However, it should not be ignored that advocacy polling exploits these lines of sensitivity to produce artificially "colored" results based upon careful question "wording" or placement.

**Conclusions**

Although we have covered the major forms of nonsampling variability that are superimposed on standard sampling error, we have not done justice to the variety of ways in which biases can potentially arise. we have seen that the conduct of a sample survey is a complex, multistage operation, with the details of the measurement process involving a lengthy series of decision points. The stakes involved in some of these decision points are not always clear. Often, however, there is no doubt about what procedures are preferable; yet choices optimal from the point of view of precision are costly in money or time.

The sum cost of procedures that improve quality but can be skipped without damage obvious to casual consumers of the survey results is large relative to the unavoidable

base costs of data collection. Thus commercial agencies can offer the needy client an eventual sample of a given size for half or even a third the normal price simply by progressively abandoning extra steps involved in a full-dress operation. Such a cut-rate edition can often be justified because client aims to not require the precision in results that may be necessary for other purposes, including scientific ones. The problem is that, of obvious reasons, these lower-quality efforts are not identified as such when results are published, making it impossible for the consumer to adjust error margins accordingly.

It should scarcely be concluded that "total error" in survey research can be reduced to mere sampling error simply by throwing money at the problem. There are various practical limits to this complex form of data collection that cannot be bought away. At the same time, it would be a pity to judge the accuracy of sample surveys from the sense of discrepancy produced by the unlabeled mix of careful and hasty results that coexist in the media. There is a broadly based conviction held by long-term practitioners that, given full quality and standardized procedures, surveys employing the same items lodged in the same context and measured at the same time will usually produce very similar results, if not within margins of pure sampling error, at least within margins that are less than half again greater. For most purposes short of "calling" very close elections, this level of precision is quite satisfactory.