Getting at the truth in Opinion Polls

By

Surjit S. Bhalla

August 10, 1991

Item 1 : May-June, 1991: In the recently concluded national elections in India, an exit poll indicated that one party (Congress plus minor allies) would get only 200 seats while its major competitor (Bharatiya Janata Party or the BJP) would get around 160 seats out of the 524 seats being contested. The final figures were Congress 244 seats and the BJP 123 seats i.e. the exit poll was considerably off the mark. This election was obviously affected by the assassination of Rajiv Gandhi, the prime ministerial candidate from the Congress party. Consequently, the opinion polls lost their significance as predictors. For the record, however, two opinion polls had the Congress getting around 240 seats and one poll had the Congress obtaining 300 seats. (See Table 1).

Item 2 : Nov, 1990: Jesse Helms is observed to be trailing Harvey Gantt, a black candidate in the election for the North Carolina Senate. The newspapers are awash with discussion that Helms, after three conservative terms in the Senate, might finally lose one ! The speculation is intense because the "accurate" opinion polls give Gantt a comfortable lead of around 8 percent in late October - only about two weeks before election day. On the day that mattered, one exit poll has the race as dead even. The pundits and pollsters are proved wrong - Helms, not only does not lose, he actually wins by a large margin of 53 to 47 percent. In other words, the gap between prediction and reality is a massive "swing" of 14 percent. The New York Times does note that " It is clear the polls overstated support for a black, as has happened in other races" (p. B3, Nov. 8, 1990).

Item 3 Feb. 1990: Nicaragua is finally holding an election and its ruler, Mr. Daniel Ortega, is expected to win with a thumping majority of 20 percent over his "unknown" rival Mrs. Violet Chamorro. When the results are out, it is a major surprise - Mrs. Chamorro is in with a large majority of 20 percent - exactly the opposite of poll predictions. In retrospect, Mr. Ortega and the pollsters

wonder as to what went wrong.

Item 4 and 5: Mr. Dinkins and Mr. Wilder - Two separate races pitch strong black Democratic candidates - Dinkins in New York (race for Mayor) and Wilder in Virginia (race for Governor). In eerie parallels, both the black candidates are favored to win by the opinion polls by comfortable margins of around 10 percent. These predictions are confirmed by the exit polls which show the candidates winning by ?? percent and ?? percent respectively. This time, the pollsters are not wrong, but embarrassed - instead of easy victories, the margins are wafer thin. (give details here).

Item 6 - Rajiv Gandhi, incumbent Prime Minister of India, in November 1989: While nobody expected him to repeat his sweeping mandate of 1984 (49.4 percent of the vote and 415 out of 545 seats in the Indian Parliament), the conventional wisdom favored him with a comfortable win of around 300 seats. Opinion polls a month before the election also echoed this expectation. However, the pollster (myself) for the prestigious Indian weekly, SUNDAY, predicted a major surprise - Rajiv Gandhi and the Congress party to win not 300 but to lose the election with a tally of only around 200 seats. In actual event, the Congress obtained 197 seats.

The gap between prediction and reality

The fact that the above "surprise" list is small (relative to the large number of opinion polls that have been conducted) suggests that these days opinion polling has become extremely sophisticated. So much so that the financial markets now pay close attention to these measured changes in public mood - the recent swings in the Conservative party's opinion poll popularity in the United Kingdom are closely related, ceteris paribus, to swings in the UK stock and bond indices. Given this record on accuracy and the consequent dependence on it of "hot money", mistakes in

polling are often deemed as major "surprises" and questions asked of pollsters - how did you go wrong? There are several possibilities of "error" between prediction and reality. First, preferences for candidates can change. Indeed, this is the most cited explanation of pollsters when their forecast goes awry. Second, a non-scientific sample may have been chosen; the sample may not represent the underlying population in terms of its composition i.e sex, income, and occupation. (This error is less likely to occur because the forecast will only go wrong if the behavior of people differs significantly according to their status - an unlikely event.) Third, non-representative constituencies may have been chosen e.g. using Princeton, rather than Trenton, as a constituency for forecasting the vote in New Jersey.

Finally, a fourth source of error occurs if the relationship between votes and seats is not properly identified. This is not generally a problem in the United States, where individual elections rarely involve more than 2 candidates. Contrastingly, in India (and several constituencies in the United Kingdom), it is the norm to have at least three candidates contesting a single seat. In such instances, projecting party strength in terms of seats rather than votes becomes a complicated exercise. For example, a two party contest involves just identifying which party can get more than 50 percent of the vote. In a three party contest, a candidate can theoretically win with a voting percentage in the wide range of 33.4 percent and 50.01 percent. Viewed differently, a landslide victory (defined as a two-thirds share of seats) in a two party game results with a vote share of around 55-60 percent. In a three party game, this vote share need be as little as 45 percent; and in a four party contest, this share need be only around 33 percent. This highly non-linear relationship between votes and seats implies that psephologists have an extra source of error. The voting percentage has to be exceedingly accurate in order for the seat prediction to be broadly correct. Consequently, pollsters devote an extra effort towards identifying homogeneous "vote" constituencies, and then

isolating the relationship between these constituencies and a state's voting pattern. It is easy to see, and understand, how errors in forecasting can occur if a wrong "representative" constituency is chosen.

Over the years, polling has come of age, at least in its scientific content. Detailed data on communities and their polling pattern has now made it extremely unlikely that a "wrong" sample is chosen. This rules out for most polling (at least in the advanced economies) the third explanation for errors reported above.

Analysis of the "determinants of voting and turnout behavior" has yielded important insights into voting preferences of individuals e.g. whether, and when men are more likely to vote than women, or blacks more than whites. Use of such information has meant that the second cause of forecasting error has also been considerably reduced.

Changes in voting preferences have often been identified as they have occurred. The most recent example of such an ex-ante swing was the Bush recovery from a 20 point deficit in June 1988 to a 5 to 10 percent lead going into the election in Nov. 1988. This rules out the first source of error.

Finally, sophisticated models have been constructed to yield accurate state level vote, and seat, forecasts from vote data from a non-random sample of select constituencies. This assumption, and reality (of scientific accuracy), rules out the fourth source of error in forecasting elections. Thus, all known explanations for the above "surprise" predictions are likely to be inappropriate. One is left with a dilemma -what explains forecasting errors? Obviously, any of the four errors outlined above may have occurred in the six surprise elections noted in the introduction. What is being asserted here is that it is unlikely that the above known sources of error led to wrong forecasts.

Wrong predictions have resulted from both opinion and exit polls. The above excuses/explanations for erroneous forecasts most often converge on the first, and most convenient

explanation - the voter changed his mind. However, it should be emphasized that this excuse cannot apply to exit polls. Such polls, by definition, are ex-post, and need be little more than simple counting exercises i.e. counts of yes vs. no in a scientifically selected sample. Since the art/science of sample, and constituency, selection is well developed there are no obvious explanations for the above documented debacle of the exit polls.

Identification of a consistent explanation for varied divergences between forecast and reality is not a simple exercise. But one such explanation does exist - the respondents to the polls lied when asked for their preferences. Even this explanation, however, needs to be assessed and differentiated from a tautology. If a forecast is wrong, then the percentage lying can be assumed to be tautologically equal to the difference in victory margins between prediction and reality (if no other sources of error are allowed).

Lying Index - Measurement

This article takes as given the accuracy of opinion polls. It offers instead a suggestion towards improving forecasts of elections where traditional assumptions may not be valid. What is offered is a consistent, and ex-ante correction, to polling data. This correction is based on an intuitively simple and age old assumption - people lie. Not all the time, and not even most of the time. Indeed, they lie very rarely. But when they do, they can cause serious errors. And given the high price of going wrong, it pays pollsters to heed, and calculate, a "lying index". If the basic premise of the lying index is accepted, the question arises : can it be measured, and measured successfully ? Success involves several dimensions, but the test should pass two in particular: it should broadly reflect lying when there is lying, and not reflect lying when there is no lying. In other words, there should be no "bias" in the estimate of lying.

Lying needs an explanation. There are three major reasons why respondents might lie, or as a Pentagon spokesperson might state, "misreveal their preferences". First, and this is more important in less developed countries (LDC's), there is the element of fear. In such countries, the polling is done person-to-person and the respondent most likely knows that the interviewer is not a local inhabitant. The dress and/or manner will reveal this abundantly, if not the accent of the interviewer. (In the most recently concluded exit poll in the Southern part of India (Andhra Pradesh) in May 1991, the magazine pollsters were mobbed, their questionnaires taken and burnt). The respondents fear is either economic, or physical. The former stems from the belief that the interviewer might be connected to the government - hence, the answers are what the voter feels the authorities want to hear. Physical fear is probably a more realistic explanation for the large amount of lying observed in India and Nicaragua. Picture the following: the interviewer stops you at a busy street corner (or the center of the village near the market) and begins asking questions. Rarely, if ever, is the respondent alone - he/she is with friends. Soon a mini-crowd gathers, a natural consequence of large underemployment and unemployment in these economies. The respondent is now "forced" to respond according to the wishes of the crowd. For example, he dare not say he is supporting party X if the crowd and/or the local powers that be (landlords, local thugs etc) are presumed to be for party Y. Hence, lying comes in by default.

The three US exit poll mishaps mentioned above are examples of lying of a third kind - embarrassment. Unless the sampling was coincidentally way off in these three polls, the explanation of the divergence is straightforward - white Democrats not wanting to admit (even anonymously!) that they were voting for a white Republican over a fellow black Democrat. Simple racism.

There is finally a fourth explanation for lying - strategic. Such lying may be used to affect the campaign of the opposition (e.g. through providing false confidence). It is unlikely to occur, but

mentioned here to complete the possibilities.

There are also two types of lying. One is that the respondent states he will vote (or has voted) for X when he actually intends to vote for Y. The other is that he will not vote for X when intending to vote for X. In a two party system, this nuance does not matter, but in a three party contest the identification of lying becomes difficult. (See table 2 below on how this important dimension adversely affected an opinion poll forecast in India).

Lying Index - Background

The methodology developed to measure lying can best be explained by recourse to the background under which it was developed. In Oct 1989, Rajiv Gandhi, the then Prime Minister, called elections a few months ahead of schedule. Opinion polls at the time agreed that Mr. Gandhi would win easily, albeit with a substantially reduced majority. The consensus forecast of seats was in the neighborhood of 300. This forecast was buttressed by two others -a private poll for the Congress party was rumored to have predicted 295 seats, and The Hindu (a national newspaper) opinion poll forecasted that Mr. Gandhi's party would win 267 seats. At the same time, the poll commissioned by the news weekly SUNDAY (and analyzed by myself) suggested that if the raw data was to be taken at face value, Mr. Gandhi's party would win by a landslide with a total vote share in the neighborhood of 55 to 60 percent, and seats above 350. (The translation of votes into seats is highly non-linear and a function of regional strength, with some 18 major states comprising the Indian electorate).

Perusal of the SUNDAY poll data showed an interesting anomaly - while Mr. Gandhi's predicted vote count was high, so were his "negatives" i.e. on various issues, a significant proportion of voters disapproved of Mr. Gandhi's handling of the economy, his handling of "communalism" or religious divisiveness, his position on regional autonomy issues pertaining to three states - Punjab,

Kashmir and Assam - and his involvement with corruption. (See Table 3). The last item had recently become an important campaign issue with accusations against Mr. Gandhi personally and his party. These data were therefore highly inconsistent with the voting pattern which suggested at face value, an extremely popular Prime Minister.

The reason the above negatives were identified is because "lying" was anticipated in the design of the questionnaire. (Most election opinion polls rarely ask respondents for their views on the economy, or their views on social policies). In addition to the (first) question on "who will you vote for if the election were held today" the respondents were asked for their views on various political and social items. In the American context, such questions would deal with questions like affirmative action, law and order, Supreme Court appointments, abortion etc.

Lying cannot be identified in a vacuum. Some assumptions have to be made regarding the views of the respondents. In particular, one important assumption is the following: respondents may (do) have a motivation for lying on the important personal question of who they intend to vote for; but on political, social, and economic issues the respondents are more at ease and do not mind stating their true views. (Actually, what is strictly needed to identify lying is the weak assumption that lying occur significantly less on questions other than the voting question). It is this difference that allows one to extract, statistically, the percentage of respondents who lie when asked about their voting preference.

Heuristically, the construction of the lying index is as follows. Assume that preferences for a candidate (party) are associated with the stance of the candidate on other issues. For example, a person who is against abortion would prefer a candidate who states that abortion should be outlawed. A person is less likely to vote for the Labor party in England if he feels that unions exert too much influence. American Jews are less likely to vote for a candidate who states that American policy

has been too partial towards the Israeli view on Palestine. American whites are less likely to vote for Dukakis if they become convinced that he is soft on crime.

However, there remains the question of a "teflon-proof" candidate e.g. Ronald Reagan. How would the lying index have worked in his case? It is clear that people did not agree with his views on several issues, yet still voted for him. Would the lying index have given a wrong answer? Most likely not. What would have happened with Ronald Reagan is that none of the determinants would have turned out to be (statistically) significant e.g. as many people for gun control voted for him as people who were against gun control. Hence, the "outliers" would likely have been few, and the analysis would have suggested that there was little, if any, lying.

As the above discussion indicates, one's voting preference may be a function of the candidate's views on various issues, and the voters socio-economic-cultural background. For different people, different factors are paramount. It nevertheless is likely that there may be (is) a well-defined statistical relationship between one's assessment of the candidate's views on various issues and one's eventual voting preference. And it is this well-defined structure that helps identify the "outliers" i.e. those people who do not fit the established pattern. For example, if I am a North Carolina white Democrat, and I feel that affirmative action has gone too far in North Carolina, and I am against abortion, then it is likely that I am lying when I state that I am going to vote for Harvey Gantt for the Senate. Now it is possible that I am not lying. To account for this possibility, the lying index nets out the outliers i.e. the percentage of people lying when they state that they are going to vote for Gantt is the difference between those who state that they are for Gantt and are likely to be not (conservative white Democrats) and those that state they are for Helms and are likely to be not (a smaller number but very liberal whites could fall into this category). The reason the net figure is taken is because statistically one would like the outliers to be on both sides of the voting spectrum.

Lying Index - A Real Life Example

Table 3 documents the views of Indian voters on Rajiv Gandhi, and his performance as Prime Minister during 1985 to 1989. The answers to the questions were categorized as positive, average, or negative. (Actual answers varied according to the question - for example, there was a yes/no answer to the question "Do you think that Rajiv Gandhi is personally corrupt"). Positive answers are given a value of 1, average answers a value of zero, and negative answers a value of -1. The maximum possible score of a respondent (on nine questions) is therefore + 9, and the minimum is -9. Tabulation of these questions according to voting intention yields a frequency distribution of answers. As expected, most (around 30-50 percent) of the answers are centered at a score of zero. It is expected that the larger the score, the larger the probability of a respondent voting for Rajiv Gandhi. Correspondingly, the lower the score (less than, say, -4) the lower the probability of voting for Rajiv Gandhi i.e the person does not like Mr. Gandhi's performance. The lying index is now simply the difference between the proportion of answers below -4 and the statement that the vote was for Mr. Gandhi minus the proportion of answers above +4 and the statement that the voter was not going to vote for Mr. Gandhi. In other words, a person who so dislikes Mr. Gandhi's views (score of -4 and below) is unlikely to vote for him, and one who so likes him (score equal to 4 or more) is very likely to vote for him. (For the statistically oriented, this simple exercise can be replicated using a probit equation which has voting intention as a zero-one dependent variable, and socio-economic characteristics, and views on the nine issues as independent variables.) Table 4 reports the lying index for selected states in the 1989 election. For many states, the lying index is above 10 percent, with an average all-India figure of 14 percent. Thus, the adjusted vote for Mr. Gandhi in 1989 became 41 percent rather than the lying induced figure of 55 percent. It is obvious that without this "correction" the SUNDAY poll would have been drastically wrong i.e it would have predicted above

350 seats, rather than the final prediction of 201 seats. As noted earlier, the actual tally was 197 seats.

The above lying index was also conducted for the state (Assembly) elections in India in February 1990, and the national election in India in 1991. Its record is impressive, but not without blemishes. While the method was exceedingly accurate in both the 1989 and 1990 elections, and for 14 out of sixteen states in the 1991 election, it was disastrously wrong for two states in 1991 (Table 2).

In the 1989 election, the lying proportion was a high 15 percent. In the State Assembly elections in 1990, the lying index was close to zero, and the forecast contained the following statement " in no state is the lying proportion more than 10 percent and the proportion is less than five percent for most states" (SUNDAY, Feb. 25, p. 61). As it turned out, the index passed the test of not signifying lying when there wasn't any with flying colors -the pattern of voting was as predicted in the 1990 Assembly elections.

The third use of the lying index occurred in the recent Indian elections. These elections were unusual in that for the first time they pitted three strong parties against each other - the middle-of-the -road Congress party led by Mr. Rajiv Gandhi, the religious right wing (but not a liberal right wing party in terms of economic policies) Bharatiya Janata Party led by Mr. L.K. Advani and the left-of-center Janata Dal party led by former Prime Minister, Mr. V.P. Singh. The lying index, and seat forecasts, were made for each of the 3 parties in 14 states i.e. 42 election forecasts. (These predictions are summarized on a zonal basis in Table 2). While the election results in some states and/or constituencies may have been affected by the assassination of Rajiv Gandhi, it is unlikely that the overall result was affected in any significant manner. The lying index was deemed to be significant for only the BJP i.e. it was predicted that the BJP percentage was higher in opinion polls than was likely to be in reality. Most election forecasts (including the Las Vegas style market), as

well as the exit poll (the only poll forecast not affected by the unexpected assassination of Rajiv Gandhi) stated that the BJP would win in the neighborhood of 255 seats. The lying index adjusted forecast of SUNDAY was for the BJP to obtain 110 seats. This prediction turned out to be the most accurate - the BJP won only 123 seats. The exit poll forecast was off by close to 33 percent.

The exit poll also predicted that the Congress party (plus allies) would get only around 205 seats. This forecast also appears to have been marred by lying - the Congress (plus allies) won 244 seats i.e. about 20 percent more than predicted. The SUNDAY opinion poll which did include lying had estimated that the Congress (plus allies) would win 302 seats - a prediction that was larger by about 25 percent.

Investigation of the state wide forecasts of the SUNDAY poll reveals that it was exceedingly accurate in 38 of the 42 forecasts. It was wrong in Uttar Pradesh where it attributed too many seats to the Congress and too few to Janata Dal, and wrong in Bihar where it committed the same error.

A post-mortem of the SUNDAY forecast, and its large pro-Congress error for two states, suggests two factors which may have been responsible. First, that these two states showed the largest proportion of lying in 1989 - 15 percent in Uttar Pradesh and 16 percent in Bihar. Further, even after adjustment for lying these two states had the largest over-prediction for seats for Congress. Thus, there seems to be a consistent pro-Congress bias in the sampling (and lying!) in these two states. Second, these two were the only states in the country with three way contests in 1991. And in such instances, the lying index may become intractable i.e. instances of I intend to vote for A, I hate B, but state that I will vote for C are very difficult to identify in a consistent fashion.

The moral of this analysis on the lying index - more work needs to be done! Especially for the relatively intractable case of three way election contests. But for more traditional two way contests, use of the suggested methodology is intuitive, realistic (people do lie) and likely to yield

rich dividends. The method also has applicability in adjusting consumption and income data (the latter is likely to be stated with error) and in the analysis of "don't knows" in opinion surveys.

Table 1

Opinion Poll Predictions - Indian Elections, 1991
(Seat forecasts for Indian Parliament)

	Congress	BJP	National Front
Forecasters:			
Opinion Polls			
SUNDAY	302	110	88
India Today	248	155	105
The Hindu	238	140	115
Exit Poll			
India Today	205	155	115
Actual Result	239	123	130

Note: All forecasts, and final result, are for the major party plus minor allies.

Table 2

Seat forecast errors, 1991 Indian election
(Data from SUNDAY opinion poll, May 1991)

	Congress		BJP		National Front	
	Pred.	Actual	Pred.	Actual	Pred.	Actual
North Zone (excl. UP & Bihar)	43	41	15	17	0	0
UP & Bihar	66	5	46	55	25	68
South Zone	108	102	5	5	17	17
East Zone	29	27	1	0	44	43
West Zone	56	56	43	44	2	0
Assam*		8		2		2
Total	302	239	110	123	88	130

Notes: (1) The forecasts are after adjustment for the lying factor.
(2) The state of Assam was not included in the forecast.
(3) Numbers may not tally because of countermanded elections in a few constituencies.

Table 3

Construction of Lying Index for the 1989 Indian Elections

Views on Rajiv Gandhi and his handling of the economy etc.	Percent of respondents: Positive	Average	Negative
Performance as PM (Prime Minister)	40	36	21
Performance vs. Expectations	25	38	28
Better PM compared to Opposition leader	59		20
Corrupt ?	69		23
Corrupt ? (Opposition leader)	37		15
Mr. Gandhi's handling of:			
Corruption situation	22	27	42
Tension between religious communities	19	24	46
Employment	19	21	51
% voting for Rajiv Gandhi	59		

Notes:(1) The answers have been coded as positive, negative, and average and should be interpreted in context. For example, on the issue of corruption a positive answer for Mr. Gandhi means that 69 percent did not think he was personally corrupt.
(2) The answers do not add up to 100 because of don't knows.

Table 4

Estimates of Lying Index (% respondents lying)-1989 election

State	% Votes for Congress Party			
	Unadjusted	Lying Index	Forecast	Actual
Andhra Pradesh	60	7	53	51
Bihar	59	16	43	28
Gujarat	40	8	32	38
Karnataka	60	9	51	49
Madhya Pradesh	71	20	51	38
Maharashtra	62	17	45	45
Rajasthan	54	6	48	37
Uttar Pradesh	59	15	44	32
West Bengal	50	10	40	42
All India	55	14	41	40