NATIONAL ELECTION STUDIES
Technical Report

Assessment of bias due to attrition and sample
selection in the NES 1989 Pilot Study

Santa Traugott and Giovanna Morchio
Center for Political Studies
March, 1990

The two waves of the 1989 Pilot Study are the latter half of a four-wave panel. Waves 1 and 2 of the Panel are the 1988 Pre- and Post Election Studies. A subsample of Post-election study respondents were reinterviewed in waves 1 and 2 of the 1989 Pilot. (Interviewing took place in Sept/Oct. and Nov/Dec. of 1988; July/August and Sept/Oct. of 1990.)

Panel data are used to assess changes over time, or change induced by change in some stimuli. In so doing, we need to be sure that we are not confounding "real" changes with changes due to panel mortality/attrition, and with bias due to sample selection. There are several sources of difference between distributions on measures in the first wave of this panel (the Pre-election study) and the last wave of the panel (wave 2 of the Pilot study).

In order to assess these differences and their possible sources, the tables below compare the distribution of the entire pre-election set of respondents on demographic characteristics and a few political variables to distributions on these same pre-election study variables for the subsets of Post-election and pilot study respondents. For example, 18% of Pre-election study respondents identified themselves as strong Democrats. (Table 1, column 1) But only 15% of the 614 respondents who eventually wound up as Wave 1 Pilot Study respondents had identified themselves in the Pre-election study as strong Democrats. (Table 1, column 5.)

The Pilot respondents are a sample of respondents. The sample universe consists of those respondents who gave us both a Pre and a Post-election interview, and who also gave us their telephone number. How biased is the sample universe of respondents, compared to the original set, i.e., Pre-election study respondents?

There is some attrition between between the Pre and Post Election Surveys. (The reinterview rate was 87 per cent.) A comparison of the first and third columns of Table 1 shows virtually no differences between the entire set of Pre-election respondents and the subset who were interviewed in the Post-election study. The largest difference is in validated vote, and it is only by 2% that post-election respondents are more likely to have voted.

Some bias may be introduced into the sample because we are using only those respondents with telephones (and who were willing to give us their telephone number). Looking at this by comparing columns 3 and 4 of table 1, there is still almost no difference in the distributions although the very modest differences that are present are consistent with what we think we know of people that have telephones as opposed to those that do not: they are of higher income, have higher levels of political information and tend to vote more.

Although the sample universe for the pilot study consists of

1

only 80% of the respondents to the pre-election survey, we conclude that the cumulative effect of Pre- to Post-election study mortality and telephone ownership is slight, with the maximum difference being that in the "universe," about 61% of the respondents can be validated as voting, compared to 56% in the Pre-election study. Other visible differences are in race and the lowest category of income.

Pilot study distributions can differ from those of the sample universe (post-election respondents who gave us a telephone number) because of a) sampling error and b) sample attrition, which is largely nonresponse. We are in the unusual and fortunate position of being able to separate out the effects of non-response and sampling error. Distributions for the 855 cases which make up the Pilot Study sample are in column 5 of Table 1, and the distributions of the Pilot study respondents, Waves 1 and 2, are in the last 4 columns of the table. Before we discuss these columns, a prefatory word about how this sample was actually drawn is necessary.

One of the main worries about panel attrition is that it occurs disproportionately among the less well informed and politically involved. This would mean that each succeeding set of respondents would, wave by wave, become progressively more biased with respect to the initial set of respondents, and specifically, they would be better informed, more interested, more likely to pay attention to politics and so on. By the end, the fourth wave would be quite "unrepresentative" of the Pre-election study respondents.

Beginning with the 1987 Pilot Study, we have attempted to prevent the gap between election study and pilot study respondents. Knowing that response rate varies by political information, we divide the sample universe into five strata according to levels of political information[1] and disproportionately draw sample from these strata according to the projected response rate of each stratum. The lower the projected response rate in a strata (the projection is based on achieved response rate by information strata in previous Pilot studies) the greater the probability of a case in that strata of being selected. (The argument is further that the achieved sample is effectively self-weighting and that weights compensating for unequal probabilities of selection are unnecessary.)

Comparing column 5 to column 4, and looking at the last part

---

[1]The political information variable is composed of measures from the Pre- and Post-election studies, including recall of candidate names, correct placement of parties on issue scales, responses to the "knowledge" questions, and the interviewer's rating of how well informed the respondent is. Another technical paper, in preparation, will discuss the construction of this variable in greater detail.

of Table 1, we note that we somewhat oversampled low information groups (26% in the sample as opposed to 20% in the universe) and undersampled the highest information strata (34% in the sample as opposed to 39% in the universe). The oversampling had very minimial effects on the rest of the distributions in Table 1, with maximum effects being on professed political interest and validated vote (naturally enough).

The oversample was intended to fix any disparity between Wave 1 Pilot Study respondents and post-election telephone respondents that is due to disproportionate nonresponse between the less well-informed and better informed. So there are two questions that we ask here: 1) how well did we did absolutely, that is, what do Wave 1 respondents look like compared to the post-election telephone respondents and 2) what might they have looked like, without the over-sample, or, how successful is the oversample gambit?

The first question requires comparisons between column 4, (Post-election telephone respondents) and column 6 (the wave 1 pilot study respondents, including the oversample). Even with the supposed nonresponse correction of the oversample in place, there are some visible differences between the sample universe and the achieved sample of the Wave 1 respondents. We have too few respondents with less than high school, too many working now respondents, too many whites, too few males, and too many voters. All of these differences are small, but they are in a consistent direction. The mortality between the two waves of the panel usually, but not always, exacerbates these trends very slightly.

It is possible to reconstruct a pilot study sample in which elements which came in as a result of oversampling are deleted. This simulated cross-section sample allows us to assess the results of oversampling, and address question 2, above. (The simulated cross-section sample is a stratified probability sample, with some efficiency gains due to stratification.)

The last two columns of the table display the simulated cross-section sample. The idea is, of course, that if we had used the cross-section instead of oversampling the less well informed, the disparity between the pilot study respondents and the sample universe would be greater. This does not turn out to be the case in a consistent fashion. Oversampling with respect to political information made us less well off with respect to gender, rural residency, and arguably with political interest. That is, while we look better in the oversample with respect to those "not much" interested in politics, we look worse with respect to those who are "very" interested. Naturally enough, since it was the basis of the oversample, we did lots better on the political information variable by using the oversample.

The case of gender is possibly instructive. The effect of oversampling on low political information is to oversample women, who are less well-informed about politics than men. Low information people tend to have lower response rates than those

3

with high information. So, oversampling low information women ought to bring us closer to a 44-56 gender distribution, since presumably otherwise women would drop out at a higher rate than men. But alarm bills should be ringing by now, because response rate is almost always thought to be lower in men than in women. So oversampling women should have the effect of driving up the proportion of women in the sample, which is just what happens. We are better off in gender terms in the cross-section because the low-information and gender variables are working at cross-purposes.

The above discussion should be set in the context of very little difference between the cross-section and the oversample. Looking just at these tables, it is hard to argue that the oversample has bought us very much, and that which we did gain may have come from the stratification itself, rather than the oversampling. (We need to think about this some more.) On the other hand, what does it cost us to use an oversample? One cost in terms of efficiency would be the requirement to use weights, but we have so far assumed that weights are not necessary.

Another concern is that we not pay too high a price in response rate, since there is a demonstrable relationship between propensity to respond to political surveys and level of political information. Looking at Table 2, we can see that the response rate for the oversample is markedly lower than for the cross-section, but the oversample is small enough so that the total response rate is not much affected. So, looking just at 1989, it's pretty hard to make a good case either for or against the oversample. It doesn't do much for us, but it doesn't cost much either.

An interesting footnote is that the relationship between response rate and levels of information is stronger in 1987 than it is in 1989. Possibly this is due to the fact that the 1989 Pilot Study wave 1 respondents already reflect two rounds of panel mortality, from Pre to the Post and from the Post to the Pilot. Most of the damage may have already been done in the Pre-to Post attrition.

The relationship between response rate and political information in 1989 would probably be stronger if we were able to create the political information variable only using measures from the pre-election study, since by using post election measures as well we implicitly factor in Pre- to Post- panel mortality. Unfortunately, it is not possible to do this, since many of the components of the political information measure are based on items asked only post election, like recall of the candidates who ran for congress. It's also possible that the political information variable as we are using it is not very well specified or constructed. We need to look at this further.

Perhaps also we need to oversample to a greater extent, to get real demographic payoff, but we would recommend against this until we resolve how the information variable interacts with

4

other variables of interest, such as gender.

Also on the agenda should be an examination of how well the oversample scheme actually did work in 1987. Presumably, if it works well in off-years, it would be desireable to keep the scheme in place for presidential years, just on the grounds of consistency.

TABLE 1

DISTRIBUTION OF 1988 AND 1989 RESPONDENTS ON PRE-ELECTION STUDY
DEMOGRAPHIC AND POLITICAL VARIABLES

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
N	2040	1854	1775	1640	855	614	494	516	420
Age (Bracketed)									
18-25 yrs	12%	12%	11%	11%	12%	10%	9%	10%	8%
26-35 yrs	25	25	25	24	25	27	28	26	26
36-45 yrs	21	21	21	22	20	21	23	22	23
46-55 yrs	12	12	13	12	12	13	13	14	14
56-65 yrs	12	12	13	13	12	12	11	11	12
66 & over	17	18	17	18	19	16	16	16	16
Summary Education									
8 grades or less	10%	9%	10%	9%	10%	7%	6%	7%	6%
9-12 (no HSD)	12	11	11	11	12	9	9	8	8
High School Diploma	36	35	35	35	36	37	37	34	33
More than 12 (no dg)	17	17	17	17	16	17	17	18	18
AA degree	6	6	6	6	5	6	6	7	7
BA degree	14	15	15	15	14	16	16	17	18
Advanced degree	6	6	7	7	7	8	9	9	10
Working Status									
Working now	63%	64%	63%	64%	65%	69%	69%	70%	69%
Temporarily Laid off	1	1	1	1	1	--	--	--	--
Unemployed	4	3	4	3	4	3	3	3	3
Retired	15	16	15	16	16	14	14	14	14
Disabled	3	2	3	2	2	1	1	1	1
Housewife	11	11	11	11	11	10	11	11	11
Student	2	2	2	2	2	1	2	2	2
Race									
White	83%	86%	84%	86%	86%	88%	89%	89%	91%
Black	13	11	12	11	11	9	9	9	8
Other	3	3	3	3	3	2	2	2	2

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
N	2040	1854	1775	1640	855	614	494	516	420
Sex									
Male	43%	43%	43%	44%	41%	40%	39%	42%	42%
Female	57	57	57	56	59	60	61	58	58
Marital Status									
Married	55%	57%	56%	57%	56%	58%	59%	58%	60%
Never Married	18	17	17	16	16	15	15	14	13
Divorced	11	11	11	11	11	12	13	11	12
Separated	4	4	5	4	4	4	3	4	3
Widowed	10	10	10	10	11	10	10	11	10
Common Law	2	2	2	2	2	2	2	2	2
Labor Union									
Union Household	19%	20%	19%	19%	18%	18%	20%	19%	21%
Non-union	81	80	81	81	82	82	80	81	79
Religion									
Protestant	65%	65%	65%	65%	64%	67%	67%	67%	67%
Catholic	24	24	24	24	24	22	22	22	22
Jew	1	2	2	2	2	2	2	2	2
Type of Community									
On a Farm	20%	20%	20%	20%	23%	23%	22%	21%	21%
In the Country	12	12	13	12	12	11	12	12	12
In a small city	32	32	32	32	32	33	32	33	32
Medium size city	10	10	10	11	10	10	10	11	12
Large city	12	12	11	12	11	11	11	10	10
Suburbs of large city	6	6	6	6	6	7	8	7	7
Very large city	5	5	5	5	4	4	3	4	4
Suburbs of very large	2	2	2	2	3	2	2	2	2

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
N	2040	1854	1775	1640	855	614	494	516	420

Occupation

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
Executive	12%	13%	13%	13%	11%	12%	12%	13%	13%
Professional	14	15	15	15	14	16	17	18	19
Technician	3	3	4	4	4	5	5	4	4
Sales	10	10	10	10	11	11	12	11	11
Admin. Support	15	16	15	16	15	15	16	14	15
Priv. Household	1	1	1	1	1	1	1	1	1
Protec. Services	1	1	1	1	1	1	1	1	1
Services except	13	12	12	11	13	11	9	10	8
Farming	3	3	4	4	4	4	5	4	4
Precision production	10	11	11	11	10	9	8	9	9
Machine Operators	9	7	8	7	8	7	7	7	7
Transportation	5	5	5	5	5	5	5	5	5
Handlers	3	3	3	2	3	3	3	3	3
Armed Forces	1	1	1	1	1	--	1	--	1

Family Income

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
Less than $7,000	13%	11%	12%	10%	12%	9%	8%	8%	7%
$7,000-11,999	11	11	11	11	12	10	9	9	9
$12,000-19,999	17	17	17	17	16	16	16	16	16
$20,000-29,999	18	19	19	20	19	20	21	19	20
$30,000-39,999	15	15	15	15	16	18	19	19	20
$40,000-59,999	15	16	16	17	16	18	17	19	18
$60,000 & over	10	11	11	11	10	10	10	11	10

Party Identification

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
Strong Democrat	18%	17%	17%	17%	17%	15%	16%	15%	15%
Weak Democrat	18	18	19	19	18	18	18	18	17
Independent Democrat	12	12	12	12	12	13	13	13	12
Independent	11	10	10	10	11	11	12	10	11
Independent Republ	13	14	13	13	14	14	15	15	11
Weak Republican	14	15	14	15	15	15	14	14	15
Strong Republican	14	14	14	15	13	14	14	15	15

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
N	2040	1854	1775	1640	855	614	494	516	420

Liberal Conservative Self-placement

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
1. Liberal	3%	2%	2%	2%	2%	2%	2%	2%	3%
2.	8	8	8	8	8	7	7	7	8
3.	13	13	13	13	12	12	12	13	13
4.	31	31	31	31	31	31	31	29	29
5.	22	22	22	22	23	23	22	22	21
6.	19	20	19	20	19	20	21	22	23
7. Conservative	4	4	4	4	4	4	4	4	4

Interest in Political Campaign

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
Very much	28%	29%	29%	30%	27%	28%	29%	31%	33%
Somewhat	47	47	47	47	47	49	49	49	48
Not much	25	24	24	23	26	23	22	20	19

Turnout

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
Yes	--	--	70%	72%	71%	74%	76%	76%	78%
No	--	--	30	28	29	26	24	24	22

Vote Validation

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
Voted	56%	59%	58%	61%	59%	65%	66%	67%	68%
Did not vote	41	38	39	36	39	33	32	31	30
DK voted	3	3	3	3	2	2	2	2	2

Political Information Index

	PRE	PR-T	POST	PO-T	SAMPLE	WAVE 1	WAVE 2	W1-C	W2-C
Stratum 1, Low Info	--	--	22%	20%	26%	22%	20%	17%	15%
Stratum 2	--	--	22	23	23	21	21	19	19
Stratum 3	--	--	18	18	18	19	18	19	19
Stratum 4	--	--	21	21	19	21	22	23	25
Stratum 5, High Info	--	--	17	18	15	18	18	21	22

The Political Information Index is built from Pre and Post variables, so it can't be constructed for the entire Pre sample.

TABLE 2

Pilot 1987, 1989 (Cross-section and Oversample) Response Rate by Political Information Index

	1987			1989		
	C-S	Ov-S	Tot	C-S	Ov-S	Tot
N	515	123	638	687	142	829
Political Info Index						
Stratum 1, Low Info N, Tot Stratum	49% 119	55% 66	51% 185	68% 134	61% 74	65% 208
Stratum 2 N, Tot Stratum	75% 108	70% 23	74% 131	65% 154	79% 33	67% 187
Stratum 3 N, Tot Stratum	81% 96	62% 29	77% 125	76% 127	73% 26	76% 153
Stratum 4 N, Tot Stratum	82% 95	80% 5	82% 100	82% 147	89% 9	82% 156
Stratum 5, High Info N, Tot Stratum	91% 97	-- --	91% 97	87% 125	-- --	87% 125
Total Sample	74%	60%	72%	75%	69%	74%