Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout

STEPHEN D. ANSOLABEHERE *Massachusetts Institute of Technology*
SHANTO IYENGAR *Stanford University*
ADAM SIMON *University of Washington*

Experiments show significant demobilizing and alienating effects of negative advertising. Although internally valid, experiments may have limited external validity. Aggregate and survey data offer two ways of providing external validation for experiments. We show that survey recall measures of advertising exposure suffer from problems of internal validity due to simultaneity and measurement error, which bias estimated effects of ad exposure. We provide valid estimates of the causal effects of ad exposure for the NES surveys using instrumental variables and find that negative advertising causes lower turnout in the NES data. We also provide a careful statistical analysis of aggregate turnout data from the 1992 Senate elections that Wattenberg and Brians (1999) recommend. These aggregate data confirm our original findings. Experiments, surveys, and aggregate data all point to the same conclusion: Negative advertising demobilizes voters.

In the December 1994 issue of the *Review* (Ansolabehere et al. 1994), we published the results of a series of experiments that showed exposure to negative advertising reduced turnout and eroded confidence in the electoral process. As a reality check, we examined the results of the 1992 U.S. Senate elections and found a similar relationship between campaign tone and turnout. In the spirit of replication, Wattenberg and Brians (1999) have examined the relationship between *recall* of campaign advertising and turnout in the 1992 and 1996 NES surveys (Miller et al. 1993; Rosenstone et al. 1996). In their logit analyses, recall of negative advertising correlated positively with turnout in 1992 but negatively, although not significantly, in 1996. Wattenberg and Brians also criticized our aggregate Senate data and presented alternative figures from the Federal Elections Commission (FEC).

External replication of experiments is essential, but internal validity is an equally important concern with aggregate and survey data. After careful statistical analysis, we find that both the aggregate and survey data corroborate our experimental findings. Wattenberg and Brians offered no analysis of the aggregate data. Had they replicated our multivariate analysis with the FEC figures, they would have found a robust, positive relationship between tone and turnout. We present that analysis here. The survey data pose a thornier set of issues. We demonstrate that recall of advertising is an unreliable indicator of actual exposure and that recall both is caused by and is itself a cause of reported turnout. Measurement error and simultaneity, we show, compromise the logit results. After correcting for these biases in the NES data, we find that negative advertising reduces turnout.

We address the survey data first before turning to the aggregate data. Because the problems we identify affect most research in political communication, we close with some reflections on the state of the literature and the standards for appropriate replication.

SURVEY ANALYSIS

The analysis by Wattenberg and Brians (1999) of the relationship between recall and turnout depends on two fundamental premises. First, recall of advertising is a close surrogate for actual exposure to advertising. Second, recall is itself not caused by turnout (it is exogenous) and is uncorrelated with any omitted predictors of turnout. Both premises are false. As we show below, the NES recall measure produces severe biases due to measurement error and simultaneity. When we correct for these biases using a valid two-stage specification, the observed effects of recall on turnout show that negative advertising indeed reduces participation.

Evidence of Bias

Guessing and errors of memory produce substantial errors in recall data (for a general review, see Bradburn, Rips, and Shevell 1987). The inaccuracy of recall has been documented across a wide range of topics, including current media use (Price and Zaller 1993), political participation (Pierce and Lovrich 1982; Silver, Anderson, and Abramson 1986), personal health (Loftus et al. 1990), and employment status (Rosenstone and Hansen 1993, 67–9).

In our own experiments, we asked participants, about thirty minutes after they had watched the experimental ad, to list the ads they could remember. Only half the participants who in fact saw a campaign ad could recall having seen one (Ansolabehere and Iyengar 1995b).[1] Clearly, recall is a poor measure of

Stephen D. Ansolabehere is Professor of Political Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307. Shanto Iyengar is Professor of Political Science and Communication, Stanford University, Stanford, CA 94305-2050. Adam Simon is Assistant Professor of Political Science, University of Washington, Seattle, WA 98195-3530.

The authors thank Alan Gerber for his useful comments.

[1] Among participants who did not watch a campaign ad (i.e., the

exposure, missing one contact for every one that it captures.

Random measurement error in independent variables biases multivariate regression estimates (see Greene 1997, 435–9). The problem may be even more severe if recall is caused by intention to vote, that is, if the relationships are simultaneous, or if important variables are omitted from the analysis. A vast literature documents the potential for simultaneity biases in surveys, which likely arise because the more politically involved are more likely to recognize, remember, and comment on political messages (Bartels 1993; Higgins, Kuiper, and Olson 1981). Elderly people, for example, are more able to remember news stories about social security, and African Americans recall stories about racial discrimination more frequently than do whites (Iyengar 1990). By extension, likely voters may have better memory for political ads. Any unmeasured factors that directly affect turnout create "simultaneity" biases, too.

Our experimental data provide a test for the possible biases in logit coefficients stemming from simultaneity and measurement errors in recall. We estimated the logit analysis reported in our original article using recall of advertising tone in place of actual advertising tone (see Ansolabehere and Iyengar 1995b). Whereas the difference in intention to vote due to actual tone was five percentage points ($p < .05$), the difference between participants who recalled a positive ad and participants who recalled a negative ad was only two points and nonsignificant. This is clear evidence that estimates of *causal* effects based on recalled exposure are biased, and the biases appear to be in the direction of the results reported by Wattenberg and Brians.

Correcting the Bias

The problems with the survey recall data are severe but remediable. As the experiments show, it is not sufficient merely to add control variables to the analysis. Rather, the solution requires untangling the variation in recall that is due to actual exposure from the variation that is due to individuals' differential abilities to remember political messages.

To do so, we treat the campaigns as natural experiments. In our laboratory experiments we manipulated exposure to advertising tone at the individual level. In the real world of political campaigns, individual-level exposure to advertising varies with the intensity of the campaign at the aggregate level. We exploit the aggregate variation in presidential campaign intensity across states and over time as a quasiexperimental analogue to our laboratory manipulations. Actual exposure to advertising in general and negative advertising in particular is considerably higher in areas where the candidates advertise more extensively and during the latter stages of the campaign. Aggregate-level indicators of campaign intensity, as we discuss below, reveal the extent to which recalled exposure reflects actual exposure.

The causal effect of exposure to negative advertising in the surveys can be estimated in two steps. First, we observe changes in recall of advertising that are attributable to actual exposure, as captured by the quasiexperimental manipulations. This becomes the "treatment" variable. Second, we measure the effect of the treatment on changes in turnout. Thus, the causal effect of advertising tone on turnout equals the change in turnout that corresponds to the change in recall produced by variation in aggregate levels of actual exposure. As demonstrated by Angrist, Imbens, and Rubin (1996), the quasiexperimental logic can be implemented statistically as two-stage, instrumental variable estimation.

The validity of two-stage estimation procedures depends on the quality of the instrumental variables (or quasiexperimental manipulations). Instruments should strongly predict recall of ads but have no appreciable direct effect on intentions to vote, except through exposure to advertising. Our instruments consist of three aggregate-level variables that affect the probability of exposure to campaign advertising. The first is *Combined Gross Ratings Points* of the presidential ad buys in each state,[2] or *Volume of Advertising*. Gross rating points measure the number of exposures to advertisements purchased by the campaigns. They provide an appropriate instrument because they determine the probability that a randomly chosen individual in a particular state is exposed to political advertisements in general and to negative advertisements in particular. Elsewhere we have demonstrated that higher volume campaigns feature disproportionately more negative appeals (Ansolabehere and Iyengar 1995a, 204–6). The same pattern holds in the NES data. Respondents interviewed in states with more total advertising reported seeing disproportionately more negative than positive advertisements (see Table 1). The second variable is *Days to the Election*, or *Date of Interview*. Wattenberg and Brians use this as a control variable to capture interest in the election. Because their regression already includes individual-level measures of interest and information, however, "days to the election" reflects cumulative exposure of the electorate to campaign messages over time. The date of the interview is an appropriate instrument for tone because campaigns become more negative as the election approaches. The NES data show just such a trend. Recall of negative advertising increased from 10% in early September to 50% in late October; recall of positive advertising increased from 7% to 17%. The third variable is *Day of the Week*. Television viewing varies systematically by day; presumably, exposure to campaign advertising varies accordingly.

To confirm the robustness and validity of our two-stage approach we estimated the causal effects of exposure to advertising in the NES data in three different ways: (1) differences in means, (2) linear

control group), 4% mistakenly stated that they had seen a political ad.

[2] We are indebted to Daron Shaw for providing these data. See Shaw 1999.

TABLE 1. Estimates of Causal Effects of Tone on Turnout (Differences of Means, 1992 and 1996 NES)

Quasiexperimental Manipulation	1992 % Intending to Vote	1992 Avg. Recall of Tone[a]	1992 N	1996 % Intending to Vote	1996 Avg. Recall of Tone	1996 N
Volume of Advertising[b]						
High ad volume	.739	−.237	840	.747	−.221	679
Low ad volume	.763	−.165	1,414	.781	−.096	855
Estimated causal effect[c] (standard error)	.33 (.19)			.27 (.13)		
Date of Interview[d]						
Late in campaign	.745	−.296	930	.757	−.190	775
Early in campaign	.767	−.052	1,320	.773	−.113	759
Estimated causal effect (standard error)	.09 (.05)			.23 (.17)		

[a] Tone is coded +1 for positive ad recall, 0 for no ad recall, and −1 for negative ad recall. Average recall ranges from −1 to +1.
[b] High and low advertising campaigns are those states in which the combined advertising buys of the Democratic and Republican presidential candidate had above-average and below-average gross ratings points, respectively. The average combined gross rating point was 6,200 (s.d. = 4,600) among the states in the 1996 NES sample and 8,600 (s.d. = 5,800) among the states in the 1992 NES sample.
[c] The formula for the estimated effect is: Turnout(High) − Turnout(Low)/Recall(High) − Recall(Low).
[d] Late campaign respondents are those interviewed fewer than 32 days before the election.

two-stage least squares (2SLS), and (3) multivariate probit. Although objections can be raised to any one of these approaches, the fact that all three produce similar results suggests that the causal effect is robust.

Differences of means provide a simple estimate of the causal effects of negative advertising in the survey data.[3] First, we divide the sample into groups corresponding to levels of each of the quasiexperimental manipulations. For the volume of advertising (gross rating points), the groups are survey respondents in states with high (above-average) levels of advertising and those in states with low (below-average) levels of advertising. For date of interview, the groups are those interviewed early and late in the campaign season. We then measure the mean level of recall and the percentage intending to vote in each group. We code recall as a trichotomy for each individual (−1 for negative ad recall, 0 for no recall, and +1 for positive ad recall).[4] Average recall equals the percentage reporting positive ad exposure minus the percentage reporting negative ad exposure.

Table 1 displays the percentage of respondents who intended to vote and the average recalled tone for the quasiexperimental manipulations. Recall of negative ads is significantly higher in states with higher levels of advertising and in the latter stages of the campaign. In addition, intentions to vote are *lower* in the states with more television advertising and in the latter stages of the campaign.[5] These facts alone contradict the conventional wisdom that campaigns necessarily increase turnout. The estimated causal effect of exposure to advertising is the ratio of the turnout and recall differences. Using volume of advertising as an instrument for tone, the causal effect of advertising tone on turnout is +.33 in 1992 and +.27 in 1996. A 1% change in the tone of the campaign (in a more positive direction) causes a .3% rise in turnout. Using time of campaign as an instrument for tone, a 1% increase in campaign tone (in a more positive direction) causes turnout to rise .09% in the 1992 survey and .23% in the 1996 survey. All these effects are in the expected direction; all but one are significantly larger than 0.

We introduce controls into the analysis using linear 2SLS and multivariate probit. Two-stage least squares uses predicted rates of advertising recall—based on all exogenous factors, including the quasiexperimental variables (interview date, gross ratings points, and day of the week)—to predict vote intentions. We use the same set of exogenous variables as Wattenberg and Brians. Imbens and Angrist (1994) show that linear 2SLS yields valid estimates of the average causal effect under fairly general conditions; those conditions appear satisfied here.[6] Because 2SLS is very inefficient, we use multiple imputations (Little and Rubin 1986) to retrieve missing data, which amount to one-quarter of the cases. The imputation procedure improves the standard errors considerably but does not change the coefficients noticeably (see Appendix Table A-1). We view the 2SLS estimates as a linear approximation of

[3] Angrist, Imbens, and Rubin (1996, 452–4) discuss the difference of means estimator and show that it provides unbiased estimates of causal effects, even without control factors.

[4] The two categories are nearly exclusive in the two surveys because almost no one reported exposure to both positive and negative advertising, although they were allowed to.

[5] The difference in recalled tone between low- and high-volume campaigns is −.214 (*t*-statistic = −10.55) in 1992 and −.125 (*t*-statistic = −4.60) in 1996. The difference in turnout between high- and low-volume campaigns is −.023 (*t*-statistic = −1.69) in 1992 and −.062 (*t*-statistic = −2.76) in 1996.

[6] The two conditions are that (1) the exclusion restrictions are valid and (2) the actual treatment is a monotonic function of the instrument (or the intention to treat, in the terminology of Angrist, Imbens, and Rubin [1996]). Monotonicity is guaranteed for the case of interview date because actual exposure is cumulative; it seems reasonable for advertising exposure and is certainly justified by the NES data. In the context of linear 2SLS we show that the exclusion restrictions are valid; see Table 2.

TABLE 2. Estimates of Causal Effects of Recall of Advertising on Turnout (2SLS and Multivariate Probit Estimates, 1992 and 1996 NES)

	1992			1996		
	2SLS Instrumental Variables for			2SLS Instrumental Variables for		
	Pos & Neg	Neg Only	Multivariate Probit[a]	Pos & Neg	Neg Only	Multivariate Probit
Recall of positive ad	.527	.017	.015	−.022	−.019	.008
	(.496)	(.022)	(.043)	(.320)	(.031)	(.035)
	−.184	−.090	−.073	−.173	−.174	−.046
Recall of negative ad	(.093)	(.049)	(.026)	(.094)	(.078)	(.018)
Number of cases	2,485	2,485	2,485	1,714	1,714	1,714
N-R^2	7.15	11.08		9.75	9.85	
[p-value]	[.62]	[.27]		[.37]	[.38]	
Durbin's test for endogeneity	6.32	19.07		5.50	8.05	
[p-value]	[.04]	[.000]		[.06]	[.004]	
F for significance of instrument set in:						
Neg ad eq.	29.85	27.72		10.03	11.05	
[p-value]	[.00]	[.00]		[.00]	[.00]	
Pos ad eq.	2.65			1.67		
[p-value]	[.03]			[.07]		

Note: Standard errors are in parentheses; *p*-values are in brackets.

[a]Probit causal effect equals the differences between the conditional probability of recall of an ad minus the conditional probability of nonrecall of that type of ad, setting exogenous variables equal to their mean values.

the average causal effect. Comparisons with nonlinear specifications, both in the single and simultaneous equation models, suggests that 2SLS approximates causal effects well even if the probability functions are nonlinear (Abadie 1998).[7] Even so, nonlinearity may create some distortion.

Multivariate probit treats positive ad recall, negative ad recall, and turnout as realizations of a multivariate normal probability function. We use the model and estimation procedure discussed in McFadden and Rudd (1994) with the additional restriction that the quasiexperimental factors do not directly influence the probability of voting. (See Appendix for details on probit specification.) Following Abadie (1998), we calculate the causal probit effect as the difference between the conditional probability of voting given exposure to a specific type of ad and the conditional probability of voting given nonexposure to that ad, correcting for selective recall.

Table 2 presents the estimated causal effects of advertising exposure from 2SLS and multivariate probit.[8] All the estimates, both linear and nonlinear, show that exposure to negative advertising lowers intentions to vote.[9] The magnitudes range from −.04 to −.17, depending on the specification; all are in the expected direction and significantly smaller than 0 at the .05 level. In no equation does exposure to positive advertising produce significant effects. This is because our set of instruments predicted recall of positive ads weakly. We reestimated the causal effects using instruments for negative advertising only (columns 2 and 5). Once again, exposure to negative advertising significantly weakens intention to vote.

An important advantage of the linear estimates is that they permit tests of the strength and validity of the instruments. At the foot of Table 2 we present the n-R^2 statistic, the F-test of the significance of the instruments, and the Durbin endogeneity test.[10] The n-R^2

[7] Evidence that the linear approximation is good comes from the comparison of logit and OLS coefficients. If the linear and logit distributions are approximately the same (i.e., nonlinearity of the vote probabilities does not distort the estimated effects), then the logit coefficients should be approximately four times larger than the OLS coefficients, as they are in these data. A further potential problem with 2SLS is that linear probability models can produce predicted values that are out of bounds; also, some of the density lies out of bounds. This problem does not appear to affect our data appreciably. Less than 3% of all cases had out-of-bound predictions; we set those cases equal to the bounds in the estimation procedure. The estimates are not affected by the constraint.

[8] The complete second-stage 2SLS estimates are shown in Appendix Table A-2.

[9] We use the conventional 2SLS standard errors to calculate the significance of the endogenous variables. Wang and Zivot (1998, 1392–3) propose a test of significance for endogenous variables when instruments in linear 2SLS are weak, as is likely the case with positive ad recall. Their test corrects for possible weakness of the instruments by using an estimate of the error variance different from conventional Wald tests. Using Wang and Zivot's test, we found that advertising exposure produced significant effects (in the expected direction) in three of the four 2SLS estimates. Only the 1992 estimates with instruments for positive and negative advertising failed, because of the unreliable estimate of positive ad recall.

[10] These tests have been developed for linear simultaneous equations with continuous endogenous variables. Our model has discrete endogenous variables. To check whether these tests still work, we ran simulations in which the statistical model involved two first-stage equations and a second-stage equation. The error distributions for all three equations were uniform. Simultaneity was created by omitting an important exogenous variable in the second-stage equation. We simulated 1,000 hypothetical data sets of 2,000 observations. We compared the tests for continuous and discrete versions of the endogenous variables. The n-R^2 and F-tests worked well, detecting failures of assumptions with nearly the same frequency in the continuous and discrete cases. The Durbin test appears to be

statistic tests the validity of the entire set of instruments.[11] In all specifications, the instrument set passes this test: The p-values are never smaller than .10. The F-test gauges the statistical power of the instrumental variables.[12] The set of instrumental variables is highly significant in the case of negative advertising; it is significant ($p < .05$) in one of the two positive advertising specifications and close to significant ($p = .07$) in the other. We, therefore, prefer the estimates that treat only recall of negative advertising as endogenous. Finally, the Durbin endogeneity test measures whether the reduction in bias achieved by 2SLS outweighs the loss of efficiency.[13] Three of the four models show that 2SLS significantly improves the estimates. The 1996 NES is a borderline case, with $p = .06$, and in this case the single-equation estimates are in the expected direction, with negative ad recall corresponding to lower turnout.

The diagnostic tests suggest that aggregate indicators of campaign intensity provide valid instruments for measuring the effects of individual-level advertising exposure on turnout. Across a range of specifications, causal estimates of the effect of advertising exposure on turnout using the NES surveys show that exposure to negative advertising significantly lowers turnout.

AGGREGATE ANALYSIS

Wattenberg and Brians (1999) argue that an alternative, and perhaps more accurate, count of the 1992 vote shows a different pattern of turnout than we reported. They recommend data published by the FEC; we obtained our data directly from the elections officers in each state in January 1993. The discrepancies between our data and those of the FEC owe largely to the treatment of absentee ballots and third parties. We originally excluded these votes; some states did not even provide us with that information. Additional differences may arise because the reports available in January 1993 contain some preliminary returns. The resulting discrepancies are idiosyncratic and unrelated to the level of turnout or the tone of the campaign.[14] As a result, these are measurement error on the *dependent* variable, which will end up in the regression error and will not bias the regression results. The proof is in the data analysis.

While our data do differ from the FEC reports, statistical analyses of the FEC data, both without and with control variables, replicate our conclusions that more positive campaigns have higher participation rates. We computed turnout (using the FEC figures, all parties and all ballots) as a percentage of the estimated voting age population (Elections Data Services 1993). Turnout averaged 60% in the positive Senate races, 54% in the mixed races, and 53% in the negative races. These figures are similar but not identical to those reported by Wattenberg and Brians (1999) in their Table 4.[15] The average rolloff was 2.6% in positive races, 5.1% in mixed races, and 4.1% in negative races.[16] The patterns of turnout and rolloff suggest that participation rates are indeed higher in states with relatively positive campaigns.

Appropriate statistical tests on the FEC data bear out our original conclusions. The F-test of the hypothesis that there is a difference in participation across the different levels of tone is 3.67 ($p < .05$). A separate F-test fails to reject the hypothesis that the effects of positive and negative campaigns are symmetric ($F = .72, p = .40$). In the case of rolloff, the pattern is more uneven and weaker than we originally found. The F-statistic for the difference in turnout rates across groups is 2.1 ($p = .13$), and that effect comes mainly from the lower rolloff in the positive campaigns.

These simple correlations may, of course, be spurious, as many factors affect turnout. In our original analysis we controlled for the closeness of the race, past turnout, education, ethnicity, income, U.S. Census form mailback rates (a measure of civic mindedness), campaign spending, open seats, and southern states. Wattenberg and Brians did not reestimate this regression, citing concerns about colinearity. These concerns are exaggerated. Multicolinearity produces inefficient estimates, which may lead researchers to accept the null hypothesis when they should not. Omitting relevant variables introduces bias. Moreover, the amount of multicolinearity here is not severe. The auxiliary regression of the tone variable on *all* the other independent variables yields an R^2 of .63, well below the level at which statisticians normally become concerned about multicolinearity.[17] It is lower still, in the range of .4 to .5, with more parsimonious statistical specifications (see Table 3).

We reproduced our original regressions using the FEC data and with the addition of a pair of dummy

somewhat biased in favor of OLS over 2SLS in the discrete data. It is, thus, a conservative test, which our specifications pass.

[11] The n-R^2 test is calculated by regressing the residuals from the second-stage regression on the included and excluded variables from the first-stage regression. If the entire instrument set is valid (i.e., all exclusion restrictions are correct), then the R^2 from the regression of the residuals on the exogenous variables should be very small. Specifically, the number of cases times R^2 is distributed Chi-square with p-j degrees of freedom, where p is the number of excluded exogenous variables and j is the number of included endogenous variables. See Greene 1997, 761–4.

[12] Several screens for instrument strength have been developed. Bound, Jaeger, and Baker (1995) discuss these, including the F-test.

[13] Several endogeneity tests have been developed. Staiger and Stock (1997) show that Durbin's test has the most power against weak instruments.

[14] The correlation between the percentage voting and the discrepancy between our measure and the FEC figures is just −0.02; the correlation with tone is −0.05.

[15] We are unsure why these differences exist. They may stem from the measure of the voting age population. We use the data published by the Elections Data Services (1993).

[16] These numbers differ slightly from the rolloff figures reported by Wattenberg and Brians because they code the 1992 Kentucky Senate race as mixed; we code it as negative. Of the 34 Senate races in 1992 that we coded, this was the only ambiguous one: The challenger spent little but ran a very negative campaign; the incumbent's campaign was mixed. All the inferences described below hold up when we omit this race from the aggregate data analysis.

[17] Johnston (1984, 247–9) shows that the parameters in an OLS regression are not sensitive to colinearity unless the R^2 in the auxiliary regression exceeds .75 or .8.

TABLE 3. Turnout and Rolloff by Tone and Control Variables (Aggregate Senate Returns, 1992)

	Turnout as % of Voting Age Population		Rolloff	
Independent Variable	Full	Parsimonious	Full	Parsimonious
Positive ad			−4.41	−4.15
			(1.25)	(.98)
Negative ad			−0.04	0.53
			(1.04)	(.92)
Tone	2.44	2.27		
(Pos-Neg)	(.98)	(.73)		
Closeness of race	−12.01	−9.62	18.78	16.61
	(8.26)	(5.71)	(4.61)	(3.83)
Mailback rate	45.67	37.44	−30.63	−29.70
	(23.56)	(14.68)	(13.49)	(8.29)
Percentage white	24.39	26.06	7.22	4.58
	(7.42)	(5.93)	(4.20)	(2.95)
1988 pres. vote as % of VAP	20.42	17.71	−1.72	
	(12.90)	(10.61)	(7.79)	
Percentage college ed.	13.58	17.45	−13.53	
	(17.95)	(10.80)	(10.24)	
Percentage over 65	−.57	−.56	−.25	
	(.35)	(.31)	(.21)	
Reg. day month before election	−4.57	−4.48	.21	
	(1.36)	(1.19)	(.77)	
Same day reg.	−.90		4.55	4.11
	(3.09)		(1.76)	(1.61)
South	.71		−2.74	−2.61
	(2.39)		(1.36)	(.92)
Open seat	.44		−1.58	−2.45
	(1.99)		(1.14)	(.78)
Per capita income (in 1,000s)	.30		.24	
	(.38)		(.21)	
Log combined spending	−.62		.10	
	(1.32)		(.75)	
Intercept	1.12	−.26	24.91	21.30
	(24.03)	(11.51)	(13.53)	(7.17)
Root MSE	3.21	2.99	1.79	1.75
R^2	.89	.89	.74	.67
F-test of symmetry [p-value]	.01 [.91]	.01 [.93]	4.94 [.03]	4.81 [.04]
Auxiliary R^2 (a measure of colinearity)	.63	.42	.63	.57
Number of cases	34	34	34	34

Note: Standard errors are in parentheses; p-values are in brackets.

variables for states with registration dates the same day as or thirty days before the election.[18] These results are presented in Table 3. As in our 1994 article (Ansolabehere et al. 1994), the tone variable is coded as a trichotomy, with values of −1 for negative campaigns, 0 for mixed campaigns, and +1 for positive campaigns. The first column of Table 3 shows the estimates with a full set of controls; the second column shows a parsimonious model, which excludes controls that an F-test revealed to have no significant effect. The third and fourth columns repeat these analyses, this time with rolloff. At the foot of the table we show the F-test for the hypothesis that positive and negative campaigns affect participation similarly.

The effects of advertising tone on Senate turnout mirror our original analysis. The coefficient for tone in the full model is 2.44 (with a t-statistic of 2.83); in the parsimonious model, the coefficient is 2.27 (with a t-statistic of 3.10). In addition, the effects are clearly symmetric. The F-statistic for the null hypothesis that the negative ad effect equals the positive ad effect has a p-value of .99 in the full model and .93 in the restricted model.[19] Substantively, these figures suggest that, on average, turnout in positive campaigns is nearly 5 percentage points higher than turnout in negative campaigns.[20]

The pattern for rolloff does differ from our earlier

[18] Wattenberg and Brians recommended these variables in an earlier version of their article.

[19] Excluding Kentucky lowers the coefficients somewhat, to 1.8 in the full model and 2.0 in the restricted model. The t-statistic for the full model is 1.81, with a p-value of .08; the t-statistic for the significance of the tone coefficient is 2.64 for the restricted model, with a p-value of .014.

[20] In positive campaigns the expected turnout is 2.4 percentage points higher than campaigns with mixed tone, and in negative campaigns the expected turnout is 2.4 percentage points lower. The differential turnout due to tone, holding other things constant, is 4.8 percentage points.

assessment. The effects of tone are asymmetric in both the full model and the parsimonious model. We thus use separate dummy variables to capture the effects of positive and negative campaigns rather than a trichotomy. The asymmetry arises because negative and mixed campaigns have about the same rate of rolloff, whereas positive campaigns have rolloff that is nearly 4% lower than other races, a highly significant drop.[21] In short, negative campaigns keep many voters from the polls. Positive campaigns have the added effect of keeping voters who go to the polls interested in offices below the top of the ballot.

CONCLUSION

Replication is vital to scientific research. We know of only one other research program, by Houston and his collaborators, that has investigated the effects of advertising tone using tightly controlled experimental methods along the lines of our study. Their findings? Exposure to negative advertising creates an "avoidance" set among viewers, which leaves them disengaged from the candidates and the political process (Houston and Roskos-Ewoldsen 1998; Houston, Doan, and Roskos-Ewoldsen 1999).

Experiments have high internal validity but need real-world confirmation. Aggregate data provide one approach. The Senate elections of 1992 clearly confirm our findings. Surveys offer another source of external validation, but measurement errors and simultaneity problems severely limit the internal validity of standard survey analyses. We have documented serious problems of internal validity with the NES recall question. These problems, which are endemic to survey research on political communication, are sufficient to explain why surveys contradict experimental and aggregate data results. In order for survey data to be taken as evidence of the causal effects of communications, researchers must present solid evidence of measurement reliability and internal validity, or they must fix the problem.

The challenge facing the field of political communication is to bridge the long-standing divide between experimental and nonexperimental methods. Instrumental variables are a natural way to correct the problems in surveys, but to date it has proved difficult to find valid instruments (Bartels 1993). Toward that end, the general method presented here—of using aggregate exposure measures to correct for biases in individual-level data on recall—holds considerable promise. For the specific problem at hand, we have presented a statistical specification that yields valid estimates of the effects of recall on turnout.

At least for the studies considered here, the experimental, survey, and aggregate data converge on the same conclusion: Negative advertising demobilizes voters.

APPENDIX: VARIABLE DEFINITIONS AND ANALYSIS EXCLUDING MISSING DATA

We follow the same coding of variables for the 1992 and 1996 NES surveys as Wattenberg and Brians. The dependent variable is *Intention to Vote or Not.* Of those coded as responding to the question, 75% of the 1996 sample and 76% of the 1992 sample stated that they intended to vote. The proportions of the overall sample that explicitly stated intention to vote were 68% in 1996 and 69% in 1992.

The endogenous variables in our 2SLS specifications are *Negative Ad Comment* and *Positive Ad Comment,* which indicate whether someone said s/he recalled seeing one of these types of political ads from the presidential campaigns. In 1992, of those who answered the question, 60% recalled neither, 12% recalled positive ads, and 32% recalled negative ads. Only 4% recalled both, and the correlation between negative ad recall and positive ad recall was .03. In 1996, of those who answered the question, 64% recalled neither, 10% recalled positive ads, and 26% recalled negative ads. Only 1% recalled both, and the correlation between negative ad recall and positive ad recall was −.10. Missing cases were treated unevenly in the NES data. In 1996, ad recall had only one missing case; in 1992, recall had 175 missing cases. We imputed values for the missing cases. One might also code nonresponse to these variables as not voting and not recalling. Doing so yields larger coefficients on ad recall in the 2SLS models.

The exogenous variables are coded exactly the same way by Wattenberg and Brians, who graciously provided us with their codes. Partisanship, for example, consists of three dummy variables: *Independent Leaners, Weak Partisans,* and *Strong Partisans.* The reference category consists of *Strong Independents.* For further description of these variables see Wattenberg and Brians (1999). The 1992 and 1996 NES surveys have markedly different incidences of missing data across different questions, which provides further justification for salvaging the missing cases where possible. In 1996, 335 cases (20% of the sample) are listwise deleted because of missing data in four variables: income, party, efficacy, and race. In 1992, 26% of the data are lost because of missing data across all the independent variables, except for age. Age contained no missing data in either subset.

We imputed values for the missing cases using the multiple imputation algorithm described by Little and Rubin (1986, 255–7) and an explicit stochastic regression model that modeled each variable for which some variables were missing as functions of other variables, as discussed in Little and Rubin (pp. 44–7, 61, 253–5). We programmed this procedure in STATA and imputed five values for each missing case. We tried an alternative imputation technique using dummy variables to indicate which observations had missing values for each variable. We included the dummy variables in the regressions as predictors. The results did not differ appreciably from those using the multiple imputations. We also constructed a multiple imputation based on the entire covariance structure of all the observed cases; again, the results were statistically indistinguishable from the multiple imputations described above. We present results with multiple imputations because they rest on more general assumptions. For the multivariate probits we set the value of the dependent variables equal to 0 if individuals did not respond to the question. We initially rounded the imputed values to 0 or 1, but Andrew Gelman of the Department of Statistics at Columbia University pointed out that this is not legitimate and might exaggerate effects of specific variables. We ran the 2SLS with the nonrespondents to the endogenous variables

[21] Excluding Kentucky does not affect this inference. Without the Kentucky case, the coefficient on positive campaigns is −3.70, with a *t*-statistic of −3.05.

TABLE A-1. Estimates of Causal Effects of Recall of Advertising on Turnout (2SLS, Missing Cases Deleted Listwise, 1992 and 1996 NES)

	1992		1996	
	Instrumental Variables for		Instrumental Variables for	
	Pos & Neg	Neg Only	Pos & Neg	Neg Only
Recall of positive ad	.513	.007	−.031	−.013
	(.448)	(.025)	(.037)	(.037)
Recall of negative ad	−.155	−.040	−.129	−.132
	(.109)	(.054)	(.099)	(.086)
Number of cases	1,837	1,837	1,373	1,373
N-R^2 [p-value]	13.34 [.21]	14.37 [.16]	11.58 [.23]	14.43 [.15]

Note: Standard errors are in parentheses; p-values are in brackets.

set equal to 0 and found that the coefficients were statistically not distinguishable from those reported.

Table A-1 parallels Table 1, except that Table A-1 corresponds to the subsample that remains once we delete observations listwise. The pattern of signs of coefficients resembles that in Table 1. In Table A-1, however, the standard errors in the 2SLS model are much larger. This is because 2SLS is very inefficient compared to OLS, and imputing the missing data helps a lot.

Table A-2 provides the complete second-stage results corresponding to the estimates in Table 2.

To specify the probit model, we assume a covariance structure for the errors in which the variances equal 1; the correlation between the two recall equations is zero; and the correlations between the vote equation and the two recall equations are to be estimated. We further assume that quasiexperimental factors do not directly enter the vote equation; see McFadden and Rudd (1994) for the formulation of the likelihood. The exclusion restrictions are very important to the estimation of the causal effects. Without such exclusions, the identification of the effects comes entirely from functional form. Also, without the exclusion restrictions, the correlations of the residual variances are much larger than those reported below. The likelihood function for the trivariate probit is very difficult to calculate, as it involves a triple integral, which we calculate using the numerical simulation SEM procedure described in McFadden and Rudd (1994, esp. 604–5), which we programmed in FORTRAN. For the 1992 estimates, the estimated correlation between the error for positive advertising equation and the error for the voting equation is nearly 0 (−.03), which suggests little selection correction. The estimated correlation between the error for the negative advertising equation and the error for the voting equation is −.33 and significantly different from 0. The log-likelihood equals −1,203. For the 1996 data, the trivariate probit produced a correlation be-

TABLE A-2. Turnout by Tone and Control Variables (2SLS, Complete Second-Stage Regressions, All Cases, 1992 and 1996 NES)

	1992		1996	
	Instrumental Variables for		Instrumental Variables for	
Independent Variable	Pos & Neg	Neg Only	Pos & Neg	Neg Only
Positive ad comment	.527 (.396)	.017 (.023)	−.022 (.319)	−.019 (.032)
Negative ad comment	−.185 (.093)	−.093 (.049)	−.173 (.094)	−.174 (.078)
Newspaper index	.002 (.001)	.002 (.001)	.001 (.001)	.001 (.001)
TV news index	.005 (.001)	.001 (.001)	.002 (.001)	.002 (.001)
Age in years	.004 (.001)	.003 (.001)	.004 (.001)	.004 (.001)
Campaign interest: Somewhat	.199 (.025)	.207 (.022)	.153 (.026)	.172 (.022)
Campaign interest: Very much	.265 (.032)	.279 (.026)	.225 (.035)	.265 (.030)
High school graduate	.180 (.028)	.164 (.023)	.102 (.030)	.102 (.030)
College	.262 (.027)	.264 (.024)	.200 (.035)	.213 (.032)
Family income	.006 (.002)	.007 (.002)	.005 (.002)	.007 (.002)
Marital status	.053 (.021)	.040 (.016)	.080 (.022)	.082 (.020)
Independent leaners	.045 (.027)	.045 (.024)	.078 (.034)	.076 (.034)
Weak partisans	.055 (.027)	.057 (.023)	.078 (.033)	.080 (.033)
Strong partisans	.126 (.031)	.131 (.026)	.169 (.034)	.175 (.034)
Political efficacy: Medium	.070 (.019)	.070 (.017)	.064 (.022)	.058 (.022)
Political efficacy: High	.052 (.023)	.061 (.020)	.064 (.024)	.068 (.024)
Race (white)	.002 (.024)	.015 (.021)	.013 (.026)	.038 (.026)
Gender (male)	.042 (.019)	.028 (.015)	.018 (.021)	.020 (.018)
South	−.079 (.019)	−.073 (.016)	−.050 (.019)	−.048 (.019)
Constant	−.077 (.061)	−.031 (.044)	−.012 (.059)	−.027 (.059)
Number of cases	2,485	2,485	1,714	1,714
F signif. of regression	42.00	52.54	31.59	32.44
Root MSE	.40	.36	.36	.36

Note: Standard errors are in parentheses.

tween the positive ad equation and the vote equation that was at the boundary (i.e., exactly 0). The correlation between the error in the negative ad equation and the vote equation is −.14 and the log-likelihood is −894.

The estimates for 1992 and 1996 were nearly identical to a bivariate probit in which reported vote and recall of a negative advertisement are endogenous and recall of a positive ad is an included exogenous variable. We view the multivariate probit as tentative, because multivariate models for discrete data are still being perfected (see Greene 1997, 911–2). Having estimated the linear and nonlinear two-stage models, it is our opinion that the linear models work fairly well as a first-order approximation to the estimates of the average causal effects. Abadie (1998) and Imbens and Angrist (1996) come to the same conclusion. We recommend that future researchers worry less about nonlinearity and more about the causal structure of data and the search for valid quasiexperiments and instruments.

REFERENCES

Abadie, Alberto. 1998. "Semiparametric Estimation of Instrumental Variable Models for Causal Effects." Department of Economics, Massachusetts Institute of Technology. Typescript.

Angrist, Joshua D., Guido Imbens, and Donald Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (2): 444–72.

Ansolabehere, Stephen, and Shanto Iyengar. 1995a. *Going Negative: How Political Advertising Shrinks and Polarizes the Electorate*. New York: Free Press.

Ansolabehere, Stephen, and Shanto Iyengar. 1995b. "Messages Forgotten: Misreporting in Surveys and the Bias Toward Minimal Effects." Department of Political Science, Massachusetts Institute of Technology. Typescript.

Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88 (December): 829–38.

Bartels, Larry. 1993. "Messages Received: The Political Impact of Media Exposure." *American Political Science Review* 87 (June): 267–85.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (June): 443–50.

Bradburn, Norman, Lance J. Rips, and Steven K. Shevell. 1987. "Answering Autobiographical Questions: The Impact of Memory and Inference in Surveys." *Science* 236 (April): 157–61.

Elections Data Services. 1993. *The Election Data Book, 1992: A Statistical Portrait of Voting in America*. Lanham, MD: Bernan.

Greene, William H. 1997. *Econometric Analysis*. Upper Saddle River, NJ: Prentice-Hall.

Higgins, E. Tory, Nicholas A. Kuiper, and James M. Olson. 1981. "Social Cognition: A Need to Get Personal." In *Social Cognition*, ed. E. Tory Higgins, C. Peter Herman, and Mark P. Zanna. Hillsdale, NJ: Lawrence Erlbaum. Pp. 395–420.

Houston, D. A., K. A. Doan, and D. R. Roskos-Ewoldsen. 1999. "Negative Political Advertising and Choice Conflict." *Journal of Experimental Psychology: Applied* 5 (March): 3–16.

Houston, D. A., and D. R. Roskos-Ewoldsen. 1998. "Cancellation and Focus Model of Choice and Preferences for Political Candidates." *Basic and Applied Social Psychology* 20 (December): 305–12.

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (March): 467–76.

Iyengar, Shanto. 1990. "Shortcuts to Political Knowledge: The Role of Selective Attention and Accessibility." In *Information and Democratic Processes*, ed. John Ferejohn and James Kuklinski. Champaign: University of Illinois Press. Pp. 160–85.

Johnston, John. 1984. *Econometric Methods*. 3d ed. New York: McGraw-Hill.

Little, Roderick, and Donald Rubin. 1986. *Statistical Analysis with Missing Data*. New York: Wiley.

Loftus, E. F., M. R. Klinger, K. D. Smith, and Judith Fiedler. 1990. "A Tale of Two Questions." *Public Opinion Quarterly* 54 (Fall): 330–45.

McFadden, Daniel, and Paul A. Rudd. 1994. "Estimation by Simulation." *Review of Economics and Statistics* 76 (November): 591–608.

Miller, Warren E., Donald Kinder, Steven J. Rosenstone, and the National Election Studies. 1993. *American National Election Study, 1992: Pre- and Post-Election Survey* [computer file] (Study #6067). Conducted by University of Michigan, Center for Political Studies. Ann Arbor: University of Michigan, Center for Political Studies/Inter-University Consortium for Political and Social Research [producers]. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor].

Pierce, John C., and Nicholas P. Lovrich. 1982. "Survey Measurement of Political Participation: Selective Effects of Recall in Petition Signing." *Social Science Quarterly* 63 (1): 164–71.

Price, Vincent, and John Zaller. 1993. "Who Gets the News?" *Public Opinion Quarterly* 57 (Summer): 133–64.

Rosenstone, Steven J., and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: Macmillan.

Rosenstone, Steven J., Donald Kinder, Warren E. Miller, and the National Election Studies. 1997. *American National Election Study, 1996: Pre- and Post-Election Survey* [computer file] (Study #6896). Conducted by University of Michigan, Center for Political Studies. Ann Arbor: University of Michigan, Center for Political Studies/Inter-University Consortium for Political and Social Research [producers]. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor].

Shaw, Daron. 1999. "The Effect of TV Ads and Candidate Appearances on Statewide Presidential Votes, 1988–96." *American Political Science Review* 93 (June): 345–62.

Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80 (June): 613–24.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (May): 557–86.

Wang, Jiahui, and Eric Zivot. 1998. "Inference on Structural Parameters in Instrumental Variables Regression with Weak Instruments." *Econometrica* 66 (November): 1389–1404.

Wattenberg, Martin P., and Craig Leonard Brians, "Negative Campaign Advertising: Demobilzer or Mobilizer?" *American Political Science Review* 93 (December): 891–9.