

## The Statistical Power of Election Studies to Detect Media Exposure Effects in Presidential Campaigns\*

John Zaller  
UCLA  
Zaller@ucla.edu

### Abstract

This paper analyzes the power of national election studies to detect the effects of exposure to political communication on vote choice in presidential elections. In particular, it examines the power of surveys (with N's from 500 to 5000) to detect exposure effects of different sizes (3, 5, 10 percentage points in aggregate) and shapes (linear, exponential, non-monotonic). The results indicate that most surveys -- and, as I suggest, many other political science datasets -- have less power than most scholars appear to think. For example, the paper finds a case in which a plausible "wrong" model that involves no exposure effect is more likely to be supported by standard datasets than a "right" model that does involve an exposure term, despite the fact that a large exposure effect is present in the population from which the data have been sampled.

9/21/00

\*Paper prepared for the 2000 Annual Meeting of the American Political Science Association, Washington, D.C. The idea for this paper originated at a conference on The Future of National Election Studies at the University of Houston in March, 1999. However, I would never have done the paper if Chris Wlezien hadn't made me. I thank Gary King for much advice (though he has not read the paper). I also thank Kathy Bawn, Kathy Cramer Walsh, Roland Strum, an anonymous reviewer, and especially Jim DeNardo for helpful comments on the paper, but stress that I alone am responsible for any errors it may still contain.

The proper size of national election surveys -- how many cases are needed to answer the questions under study -- has not been the subject of much analytical effort. In recent elections, the *grand dame* of election surveys, the American National Election Studies, has tapped roughly 1,800 respondents in its pre-election wave. Some national election studies have gone as low as 1,000 respondents and several European election studies now routinely survey 3,000 to 5,000 persons. In the current election cycle, the Annenberg School for Communication is setting a record for academic campaign studies by conducting some 100,000 interviews. But one searches in vain for analytical justification of these choices. At least in the United States, scholars seem to decide the size of election studies by holding a finger to the wind and asking how much funding agencies are willing to spend.<sup>1</sup>

The scholars who analyze these surveys do no better. To judge from most published work, whatever survey they are analyzing is plenty big enough to answer the question they are asking. If they get a null result, the normal assumption is that the hypothesis under investigation is weak or wrong.<sup>2</sup>

This state of affairs is odd. It is as if astronomers built telescopes and set out looking for new celestial objects without first calculating what can be seen through a telescope of given resolution. The primary suggestion of this paper is therefore that both those who create and those who consume election studies should be more self-conscious about what their measuring instrument can accomplish. In technical jargon, they should engage in statistical power analysis.

Unfortunately, power analysis in the domain of electoral research, and probably other non-experimental domains, is not straightforward. To show how it can be done, this paper reports an extended analysis of statistical power for one important research problem, media exposure effects in U.S. presidential election campaigns. For reasons that will become apparent, the paper relies on simulation rather than closed-form analysis. It reports a series of Monte Carlo experiments to determine the statistical power of campaign surveys of given size to determine whether persons who are more heavily exposed to the mass media are more likely to register the effects of political communication in their vote decisions. The method is to create a "campaign exposure effect" in realistically simulated data and determine how frequently statistical analysis can recover evidence of it in datasets of different sizes.

The results of the study indicate that the vast majority of election studies lack the statistical power to detect exposure effects for the three, five, or perhaps ten point shifts in vote choice that are the biggest that may be reasonably expected to occur. Even a study with as many as 100,000 interviews could, depending on how it deploys resources around particular campaign events, have difficulty pinning down important communication effects of the size and nature that are likely to occur.

To be sure, this finding applies only to a single research problem. Other problems have other data needs and existing election studies are sufficient for many of them. Yet the results of this paper show

---

<sup>1</sup> This statement is based on my experience serving two terms on the Board of Overseers of the National Election Studies and on occasional service as a consultant to other surveys.

that power cannot be taken for granted for any problem in which scholars are interested, even when the sample size is large.

Accordingly, this paper has two aims. The first is to provide an analytical basis for justification of the appropriate size of election studies. The hope is that, once such a justification has been more fully developed, it will persuade funding agencies of the need for larger surveys for at least some kinds of problems. The second is to show how to determine what can and cannot be reliably accomplished with existing datasets. The hope here is that scholars who turn up null or trace findings for effects "that just have to be there" may hereafter accompany their null reports with statements of the power of their data and model to detect the effects they expected to find.

Although this paper focuses on the problem of detecting exposure effects in campaign surveys, the need for attention to power analysis is not notably less acute in comparative politics, international relations, or other branches of American politics. In the course of preparing this paper, I was surprised to discover that even some experimentalists have only passing familiarity with power analysis. Thus, the exercise undertaken in this paper may have wide significance.

#### ANALYSIS VS. SIMULATION

One could, in principle, analyze the power of a survey to detect communication effects by formal analysis or by simulation. In the analytical mode, one uses a statistical formula to determine the probability that a survey of given size can detect an effect of given size. In simulation, one generates artificial datasets that embody particular effects and then tests how often the simulated effects can be recovered in data analysis.

Suppose, for example, that one wants to know the standard error of a simple random survey whose sample mean was  $p = .40$  and whose size was  $n = 1,000$ . To solve this problem by simulation, one would obtain a coin known to come up heads 40 percent of the time -- or, more likely, the computer equivalent thereof -- and flip it over and over in sets of 1,000 tosses each. Each sample proportion  $p$  from each set of 1,000 tosses would be close to the coin's true value of .40, but there would be small variation in the  $p$ 's from one sample of tosses to another. Taking the SD of the  $p$ 's in these samples would be a perfectly valid method of estimating the SE of an actual random survey with sample  $p = .4$  and  $n = 1,000$ .

In practice, of course, one would never estimate the SE of a poll by simulation, since a simple analytical formula exists to do the job much more easily. But simulation could do the job to an arbitrarily high level precision.

In this paper, I opt for simulation. Formulae exist to calculate the sample size needed to test relationships between variables in regression analysis (Cohen, 1988; King, Keohane, and Verba, 1997, p. 213), but they have important limitations. They do not formally incorporate the effects of measurement error, nor do they accommodate the residual error that arises when, as in American voting studies, the

---

<sup>2</sup> I do not exclude myself from this generalization.

dependent variable is a dichotomy. Hsieh, Block, and Larsen (1998) present a power formula for logit models, but it is limited to cases in which the independent variable is either dichotomous or normal. Since, as we shall see, the independent variables that one must use in the study of communication effects have markedly non-normal distributions, this formula has little value in the present context.<sup>3</sup> Excellent programs exist for doing power calculations for many types of problems, but they are based on existing formulae and so share their limitations (e.g., NCSC 2000<sup>4</sup>). Thus, simulation is the best available tool for assessing the power of surveys to detect the effect of media exposure on vote choice.

#### ISSUES IN SIMULATION

The paramount issue in simulation is creating data that match relevant characteristics of actual data. In the matter of flipping coins to simulate random sampling of 0-1 outcomes, this is easy to accomplish, because using a computer to flip coins is reasonably close to how random surveys are done in practice and perfectly equivalent to how random surveys are supposed to be done. (Insofar as actual survey sampling departs from purely random sampling, the analytical calculation of SEs will be just as far off from the truth as the coin-flipping simulation I have described.)

By contrast, in the matter of simulating a campaign effect in an election survey, there are numerous ways in which simulated data might fail to resemble real data. I shall survey these danger points in detail in a moment. Meanwhile, I offer this bare-bones introduction to how I have gone about doing the simulations in this paper.

1) *Specify an interesting and plausible campaign effect.* In the 1992 fall campaign, for example, Ross Perot lost about five percentage points in the polls after the media heaped scornful coverage on him for his statement that he had temporarily quit the race in July because of fear that Republicans would sabotage his daughter's wedding. As other evidence shows, it sometimes occurs that people are influenced by political communication in proportion to their level of exposure to it. It is therefore plausible to posit for analysis the following general campaign effect: News coverage of a candidate gaff causes a five-point drop in the polls for the given candidate, and the amount of the fall-off is a linear function of self-reported degree of exposure to the news.

2. *Generate artificial but realistic datasets of public opinion data that embody the campaign effect that has been chosen for study.* The datasets must resemble actual opinion data in the elementary sense that they consist of hundreds or perhaps thousands of cases, where cases are understood as "individuals" with scores on variables such as partisanship, media exposure, vote choice, and time of interview (i.e., before or after the event that is supposed to have caused a campaign effect). The simulated data are sufficiently realistic that a naïve analyst could mistake them for actual survey data.

---

<sup>3</sup> In the analyses below, the variable of interest will be the interaction between a 0-1 variable (for pre- and post- campaign event) and a skewed media exposure variable. The distribution of this interaction variable resembles no standard distribution.

<sup>4</sup> The program is available for downloading at <http://www.icw.com/ncss/pass.html>.

Further, the simulated data are constructed such that, as in actual survey data, individuals' partisanship and demographics explain most of the variance in vote choice, but that some other variable of interest, such as media exposure, explains a small amount of variance in the vote after the occurrence of the campaign event -- enough variance so that, as in the Perot example, aggregate exposure to news in the dataset as a whole creates a five-point dip in candidate support. Thus, people heavily exposed to the news might undergo a 10 percentage point drop and those with low exposure no drop at all, for an aggregate drop of five percent. Finally, the data are constructed so as to include random measurement error, multicollinearity among independent variables, and other features of actual survey data.

3. *Using hundreds of artificial datasets of given size, determine how often an appropriate regression model finds statistically significant evidence of the specified campaign effect.* In the Perot example, we might specify a regression in which support for Perot (a 0-1 variable) is the dependent variable and "Media Exposure X Time of Interview" (an interaction term) is one of the independent variables. "Time of Interview" would be coded as 0 before the campaign gaff and 1 afterwards; "Media exposure" would be a self-report of attention to the mass media. The coefficient on "Media Exposure X Time of Interview" would then indicate whether, as would be expected, Media Exposure has more effect on support for Perot after the gaff than before. Due to sampling variability, the regression would not always turn up a statistically significant coefficient for "Media Exposure X Time of Interview", but it would do so more often when samples were larger. The question is: How large must samples be in order for an appropriate regression model to reliably capture a campaign effect that is actually in the data (along with sampling and other forms of random error).

#### SIMULATING CAMPAIGN DATA

The key to doing a valid simulation of this kind is creating data that match all relevant characteristics of actual data. But what characteristics are relevant? And how do we determine numerical values for these characteristics?

These are the issues taken up in this section, which has three parts. First, I survey literature on campaign effects and specify the particular effects I wish to analyze in my power simulations. The idea is to make the campaign effects I simulate -- that is, the dependent variable in my simulations -- comparable to campaign effects that typically occur in the real world. Second, arguing from statistical theory, I develop the general characteristics of the independent variables that must be matched in the simulation. For example, I show that it is necessary in power simulation to match the variance of one's simulated independent variable to the variance of the independent variable one would use in a real world study. Finally, I use National Election Study data to determine the particular numerical values of these general characteristics (e.g., the variance that my simulated media exposure scale must have). Upon the completion of these steps, the simulation, data analysis, and presentation of results go quickly and easily.

Specifying the campaign effects to be studied

Campaign effects come in many shapes and sizes. In this paper, I simulate cases in which an important campaign event occurs at a discrete point in the campaign, and in which individuals are affected in relation to their degree of media exposure. There are, admittedly, many other kinds of campaign effects that can occur. Two particularly important ones are priming effects that cause a shift in mean support for one of the candidates, and lagged effects that develop gradually over a period of days rather than all at once. Moreover, the different types may blend into one another -- for example, a priming effect that develops over time among those most heavily exposed to the news. In this initial foray into simulation of campaign effects, I limit myself to relatively straightforward exposure effects of the kind indicated. Nor do I deal with the knotty question of the persistence of campaign effects.<sup>5</sup>

There are two independent dimensions to exposure effects. One is magnitude: What percent of the public is affected? The other is shape or incidence: Which people at which levels of media exposure are affected? I deal with each in turn.

A good compendium of the magnitude of recent campaign effects -- none of which is linked to individual differences in exposure to campaign stimuli -- is found in Holbrook's (1997) Do Campaigns Matter? The most reliable effect in presidential campaigns is the "convention bounce," the gain in straw polls that a candidate picks up in the aftermath of a nominating convention. According to Holbrook's survey of data from 1984 to 1992, the mean convention bounce was 6.5 percentage points with an SD of 3.75 (Table 4.1). Examining data for the period 1952 to 1992, Shaw (1999) found a mean effect of 7.4 points with an SD of 3.3. From my calculations, convention bounces in 1996 and 2000 were probably at the long-term average.

Presidential debates are another predictable campaign effect, but a generally smaller one. According to Holbrook, the mean of the six presidential debates from 1984 to 1992 was 2.2 with an SD of 1.3. The most a Democrat or Republican gained was 3.5 percentage points, but Ross Perot in 1992 may have gained close to five points in one of the debates, according to Table 5.4.<sup>6</sup> Shaw finds that the immediate effect of debates averages a little less than two percentage points, but that the effect tends to grow over time. Again, there are no links to campaign exposure.

In their outstanding study of the 1988 Canadian election, Johnston, Blais, Brady and Crete (1992) argue that the one of the candidate debates was the pivot of the entire election, generating a 10 point bounce for one party, priming attitudes toward a key issue, and setting the agenda for the rest of the campaign. These authors do link campaign effects to a measure of campaign exposure, but there is no test of the statistical significance of the exposure effect and little emphasis on the finding. Although the

---

<sup>5</sup> Shaw (1999) finds that the effects of conventions and debates tend to persist or increase over time, but that the effects of gaffes tend to spike and disappear. However, there is likely a fair amount of heterogeneity within the categories of Shaw's analysis that the best generalization is probably "sometimes campaign effects persist and sometimes they don't."

reason for underplaying exposure effects in a study of campaign dynamics is not made explicit, it seems obvious enough: Lack of power.

Dramatic media stories also sometimes occur, though not at predictable times. By Shaw's (1999) coding, candidates and the media cooperate to create an average of about one major candidate "blunder" per election, with the blunder rate increasing over time. An example is Ronald Reagan's declaration in August 1980 that the country was in a "depression." Perot's remark about his daughter's wedding would also meet Shaw's coding criteria. Shaw finds that blunders cost candidates an average of 6.3 percentage points.

In light of these data, the magnitudes I examine are mean shifts in the vote of three percent, five percent, and ten percent. Campaign effects much bigger than this are unusual; effects much smaller turn out to be too small to be detected except under unusual circumstances. The values I have chosen correspond to something like a big presidential debate effect, a typical blunder, and a large convention bounce, respectively.

Let me turn now to the shape of campaign exposure effects. The three most obvious possibilities are that those most exposed to the mass media may be most affected, that those most exposed to the media could be least affected, and that people at middle levels of media exposure could be the most affected. For purposes of power calculations, the first and third of these possibilities can be regarded as identical linear effects (with opposite signs).

Another important possibility is what I shall call the "elite exposure effect." This is an effect in which a campaign event affects no one in the bottom half of the attentive public, and affects people in the upper half of the attentiveness spectrum in proportion to level of media exposure. This effect is an *elite* effect because it affects only the upper part of the electorate, and an *exposure* effect because it is proportional to exposure.

Figure 1 presents an example of an elite exposure effect. The figure shows public opinion just before and just after the media frenzy over Perot's statement about his daughter's wedding. If we accept education as a proxy for media exposure, we see that the pattern conforms to the conditions of the elite exposure effect -- no change among persons moderate-to-low on exposure, and changes that are linear with respect to exposure among the rest. (Direct measures of media exposure are unavailable in this dataset.)

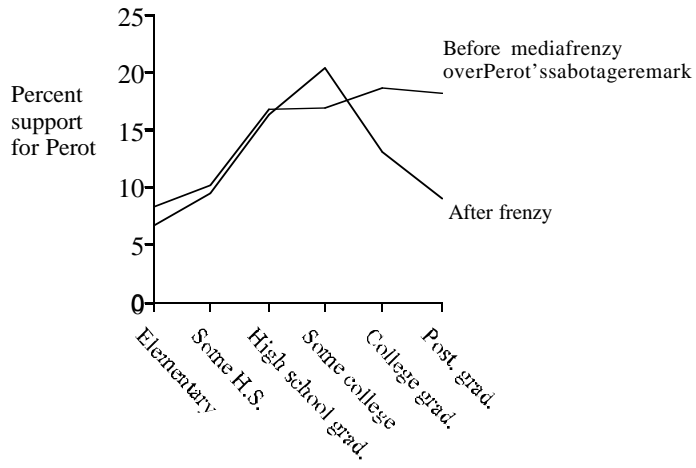
INSERT FIGURE 1 ABOUT HERE

Figure 2 shows the shape and magnitude of the particular exposure effects I have just described. A few comments are in order.

---

<sup>6</sup> However, Holbrook's Figure 5.3 makes it appear dubious that debates were wholly responsible for these gains. From tracking poll data I have analyzed, Perot's average gain was 1.2 points per debate (Hunt and Zaller, 1995).

Figure 1. *Education and responsiveness to Perot campaign story*



Source: ABC News-Washington Post tracking polls

INSERT FIGURE 2 ABOUT HERE

Note, first of all, that there is a linear positive relationship between exposure and support for Candidate A in the baseline condition in all panels in Figure 2. This reflects the fact that this zero-order relationship exists in actual data. In my simulation, as in the actual data, the correlation is spurious and has no direct effect on the vote once controls are imposed. But it is nonetheless important, for reasons explained below, to incorporate the correlation into the simulations, as shown in Figure 2.

In the data analysis section of this paper, I use a series of logit models to recover the campaign effects built into the simulations. The first of these models, shown below, will be used to recover linear exposure effects of the kind shown in the top panels of Figure 1:

$$Vote_i = b_0 + b_1 Exposure_i + b_2 Time_i + b_3 Exposure_i \times Time_i + b_4 X_{i4} \dots b_k X_{ik} \quad eq. 1$$

Exposure in this equation is an individual-level variable standing for level of exposure to the mass media. The specification and construction of this variable are described at length in the next section. The Time variable is a dichotomy, taking the value of 0 before the campaign event and 1 afterwards. Vote is also a dichotomous variable, taking the value 1 if the individual supports Candidate A and 0 otherwise. The variables  $X_4$  to  $X_k$  are control variables, as described in the next section. The form of the statistical model is the same as used to create the campaign effect in the simulation. Given all this, my test of statistical power will be how often I am able to recover a statistically significant  $b_3$  coefficient when I run this model on the simulated data.

For the non-monotonic effects, I use the following logit model:

$$Vote_i = b_0 + b_1 Exposure_i + b_2 Time_i + b_3 Exposure_i \times Time_i + b_4 Exposure_i^2 + b_5 Exposure_i^2 \times Time_i + b_6 X_{4i} \dots \quad eq. 2$$

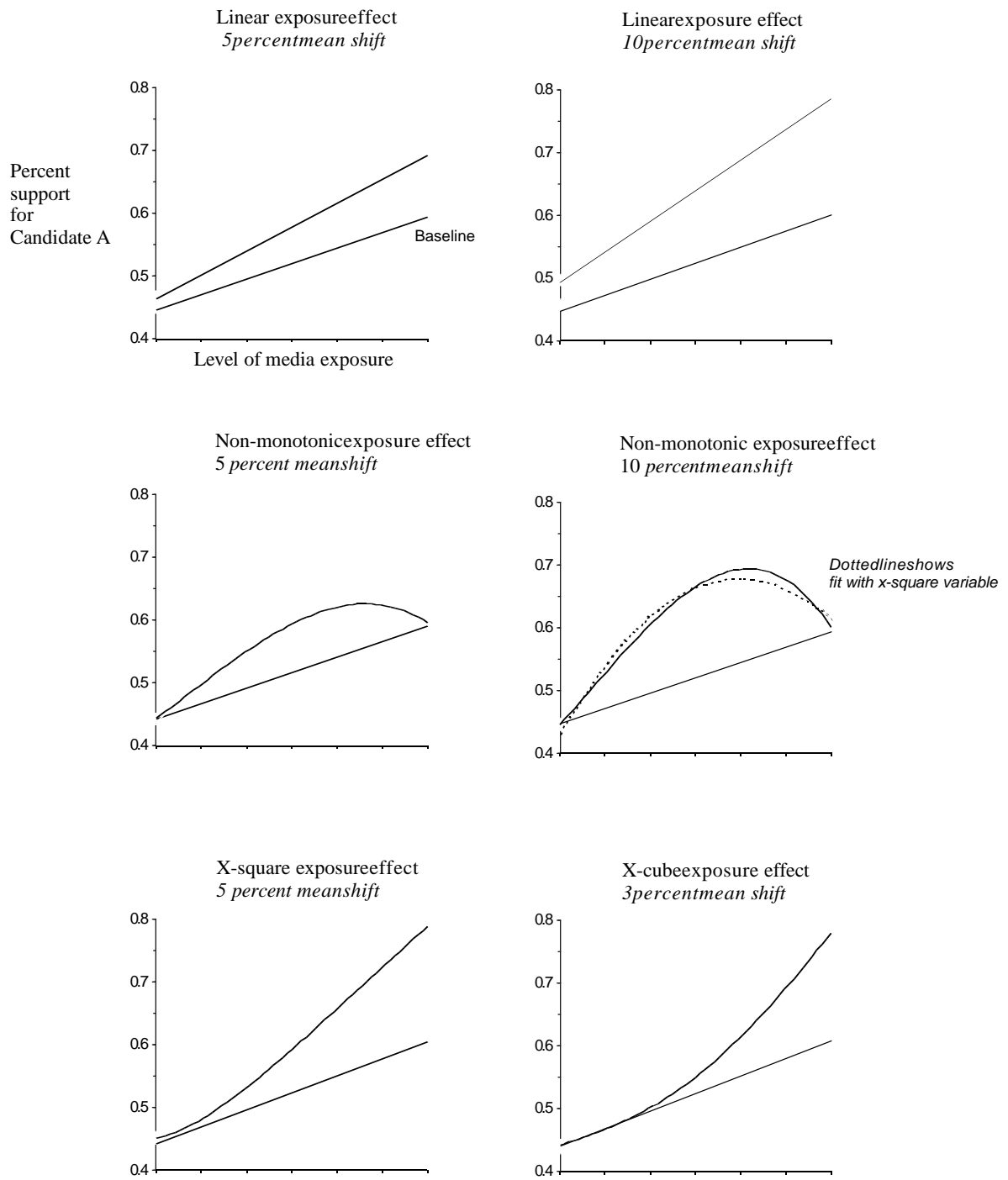
The coefficient of critical interest here is  $b_5$ , since this is the coefficient that determines the downturn in persuasion at high levels of media exposure. Unless this coefficient is both negative and achieves statistical significance, there is no clear detection of a non-monotonic effect.

This model, I must emphasize, is similar but not identical to the model used to create the campaign effect in the simulation. The difference is that, in simulating the campaign effect, I used a media-cubed variable rather than a media-squared variable. I used the media-cubed variable because it was, as I found, necessary to achieve the shape I think I see in real data; and I used the media-squared variable to test for the campaign effect because this is what I believe researchers would typically use to test for non-monotonicity.<sup>7</sup> Note that the figure for the 10 percent non-monotonic effect has a solid and a dotted line. The solid line shows the actual data, as created from an equation with an x-cubed term; the dotted line shows a fit with an x-squared term. I discuss this matter further below.

---

<sup>7</sup> More complicated models, such as those used in Zaller (1992), are beyond what I can do in this paper.

Figure 2. *The size and shape of campaign effects to be simulated*



The models used to test for the elite exposure effect are the same as actually used to simulate the effects. For the five percent and three percent effects, the models are, respectively:

$$Vote_i = b_0 + b_1 Exposure_i^2 + b_2 Time_i + b_3 Exposure_i^2 \times Time_i + b_4 X_{4i} \dots \text{ eq. 3}$$

$$Vote_i = b_0 + b_1 Exposure_i^3 + b_2 Time_i + b_3 Exposure_i^3 \times Time_i + b_4 X_{4i} \dots \text{ eq. 4}$$

The reason for using the media-cubed term in the three-percent model is that I couldn't otherwise achieve the sort of elite effect I was trying to simulate. In particular, I couldn't mimic the pattern in Figure 1 without cubing the media variable.<sup>8</sup> In the analysis that follows, I shall also test the ability of a simple linear model to capture these two media effects, since I think that many analysis would, in this situation, prefer a conventional linear model to an apparently *ad hoc* x-square or x-cube model. The reason for not testing 10 percent variants of the elite exposure models is that they would require unrealistically high levels of opinion change among the most attentive.

A final point that may seem technical but is actually quite substantively pregnant is as follows: For the two linear models in the upper panels of Figure 1, support for Candidate A is higher at  $T_1$  at every level of exposure, even when exposure is at zero. Thus, the y-intercept is slightly positive rather than, as might have been expected, zero. The slightly positive intercept captures the idea that a few people who have zero media exposure may nonetheless respond to campaign events, presumably because they discuss politics with people who have higher levels of media exposure (Huckfeldt et al.).

This specification effectively reduces campaign effects of 5 percent and 10 percent thus to media exposure effects of roughly 4 percent and 8 percent, with the rest of the campaign effect occurring by word-of-mouth and getting absorbed by the intercept. This, in turn, makes it harder to recover statistically significant exposure effects. This feature of my simulations seems to me clearly realistic, but the reader should keep in mind its effect on power.

The reader should keep in mind that my tests for the non-monotonic and elite exposure effects do not include "discussion effects" (i.e., a slightly positive intercept). The reader should likewise remain aware that I attempt to recover linear effects with the "right" model but use a slightly mis-specified one for the non-monotonic effects. There are reasons for these choices. I thought that discussion effects would be more likely to occur for a linear effect than for an elite effect which, as conceptualized, does not penetrate to the lowest strata of society. Ditto for non-monotonic exposure effects, which are often defined by the idea that the persuasive message does not penetrate all strata of society (Converse, 1962). I also thought that researchers would be more likely to use the "right model" when they happened to be analyzing a linear effect than when analyzing a non-monotonic one.

---

<sup>8</sup> I omitted first order exposure terms from these models because I happened to know that they did not belong in the "true model." Including them would have led to multicollinearity and hence reduced statistical power. Although it is unusual to use second and third order terms without lower order ones, it is not any different, in principle, from using log transformations without lower order terms.

## General issues in matching simulated to actual data

Having specified the campaign effects I wish to analyze, I now turn to the question of how to embed these effects in simulated datasets that match actual datasets in "relevant respects." What are relevant here are any characteristics of actual data that affect statistical power. I shall therefore organize this discussion in terms of threats to statistical power.

Since power increases as standard errors decrease, the discussion begins with the formula for the standard error (SE) of an OLS regression coefficient. By analyzing this formula, we can determine several of the most important threats to power.

The formula for the SE of the coefficient for variable X in a multiple regression having k independent variables may be written in standard notation as:

$$SE_{slope\ of\ X} = \left[ \frac{\sum_i^n \frac{e_i^2}{(n-k-1)}}{(1-R_{aux}^2)(n * Var(x))} \right]^{1/2}$$

I shall now translate this off-putting formula into English and show its implications for power analysis. There are four points, each keyed to a particular part of the formula.

- More cases make for smaller standard errors. The intuition here is that more cases mean more information and hence less uncertainty (i.e., smaller SEs) about the coefficient. As regards the formula, note that the denominator includes n, so that as n increases, the SE decreases (all else equal).
- Increasing the range or "spread out-ness" of the X-variable makes for smaller standard errors. The intuition here is best conveyed by example: One can more precisely estimate the impact of education on income from a sample in which educational attainment varies from first grade to Ph.D. than from a sample in which education varies from some college to Ph.D. As regards the formula, if the values of the X-variable are spread out over a large range (rather than bunched together in the middle of a range), the variance of X will be greater, which will increase the size of the denominator, thereby decreasing the standard error.
- Less residual (or unexplained) variation makes for smaller standard errors. The intuition is that one can more precisely tell what is going on in a dataset when the random noise in it is low. As regards the formula, the numerator in the formula is the variance of the residuals of the regression. If, all else equal, the variance of the residuals is lower, the numerator of the above fraction will also be lower, thereby reducing the SE of the coefficient and increasing statistical power.

As is well-known, smaller residuals also make for a higher r-square. Thus we can also say that higher r-squares make for smaller standard errors and more statistical power. (The r-square to which I refer in

the previous sentence is not the r-square that appears in the formula; that r-square is discussed in the next bullet.)

- Multicollinearity reduces statistical power. Multicollinearity is defined as the correlation between a given X-variable and the other independent variables. As is widely known, a high level of multicollinearity causes higher standard errors and hence lower levels of statistical power. The intuition is that a high degree of correlation among the independent variables makes it hard to tell which are important and which are not, which is then reflected in large standard errors.

In the formula, multicollinearity is measured by the r-square of an auxiliary regression in which the X-variable of interest is regressed on all other independent variables in the main regression are the predictor variables. The higher the r-square, the greater the multicollinearity. Since the formula for the SE contains the term  $(1 - R_{aux}^2)$  in the denominator, high multicollinearity increases the SE which in turn reduces statistical power.

Multicollinearity is an especially serious problem in the detection of campaign communication effects, for this reason: As modeled at the individual level, campaign effects typically involve a campaign-induced change in the effect of key independent variables on vote choice. Such changes are naturally modeled in terms of interactions. For example, if an effect of a campaign is to prime a certain variable so as to increase its impact on vote choice, then the effect of that variable is appropriately modeled as an interaction with time. The particular campaign effects described in the previous section also involve interaction terms, as noted.

To estimate interactions between one variable and another, one must include each variable twice -- once for its direct effect and once for its interactive effect. These two variables will be at least moderately correlated, which creates the problem of multicollinearity. Models to capture non-monotonic effects are doubly cursed: terms for Exposure and Exposure-squared, plus the interaction of Time with each. Thus, the exposure variable appears four times in the model (see Eq. 2 above).

Interaction effects are not the only source of multicollinearity. Consider the priming effect documented by Johnston et al. in the 1988 Canadian election. In that campaign, the candidate debate, news coverage, and advertising all function to prime the effect of attitudes toward a Free Trade Agreement (FTA) on vote choice. But trade policy had been the subject of partisan contention for some time and so was likely to be correlated with party affiliation and other attitudes, thereby presenting an issue of multicollinearity. This sort of problem is common, in that many political attitudes -- on party, race, abortion, defense spending, government services -- are likely to be at least moderately correlated with one another. It will therefore be important to build an appropriate amount of multicollinearity into the simulations.

The latter three bullets are of the greatest importance for simulation. If the variance of the X-variable, the size of the residuals, and the degree of multicollinearity as measured by an auxiliary r-square

determine statistical power, then we shall need in our simulations to match these entities to what typically occur in actual data.

A final theoretical issue is measurement error. No term for measurement error appears in the formula for the SE. This is because the standard SE formula assumes that all variables have been measured without error. But measurement error is nonetheless universally important in survey work, causing biased coefficient estimates and reduced statistical power. Hence, building an appropriate form of measurement error into simulations adds to the validity of the simulations.

#### Determining the empirical values for variables in the simulation

In the previous section we saw that simulated data must match actual data in particular ways. The next step is to determine the empirical values that must be matched. Given, for example, that simulated data must have the same amount of measurement error as actual data, we must determine how much measurement error exists in actual data -- and, in particular, how much measurement error exists in the media exposure scales that are typically used in studies of campaign effects. For each of the characteristics on which our simulated data must match actual data, we must determine what actual election study data are like.

The exposure variable. The central variable in this analysis, media exposure, is also the most challenging to simulate in a credible manner. The 1996 NES study offers several possibilities for measuring it, but none is ideal, and scholars have achieved no consensus on which is best. The lack of agreement is not for want of effort. The literature is full of conflicting advice and argument on how to "measure media exposure right." Meanwhile, different measures have different variances, reliabilities and auxiliary correlations with other variables, each of which characteristics affects the power of statistical analyses in which the media variable is used. In these circumstances, there is no avoiding a review of the terrain. I shall keep this review both empirical and tightly tethered to the options available in the data at hand.<sup>9</sup>

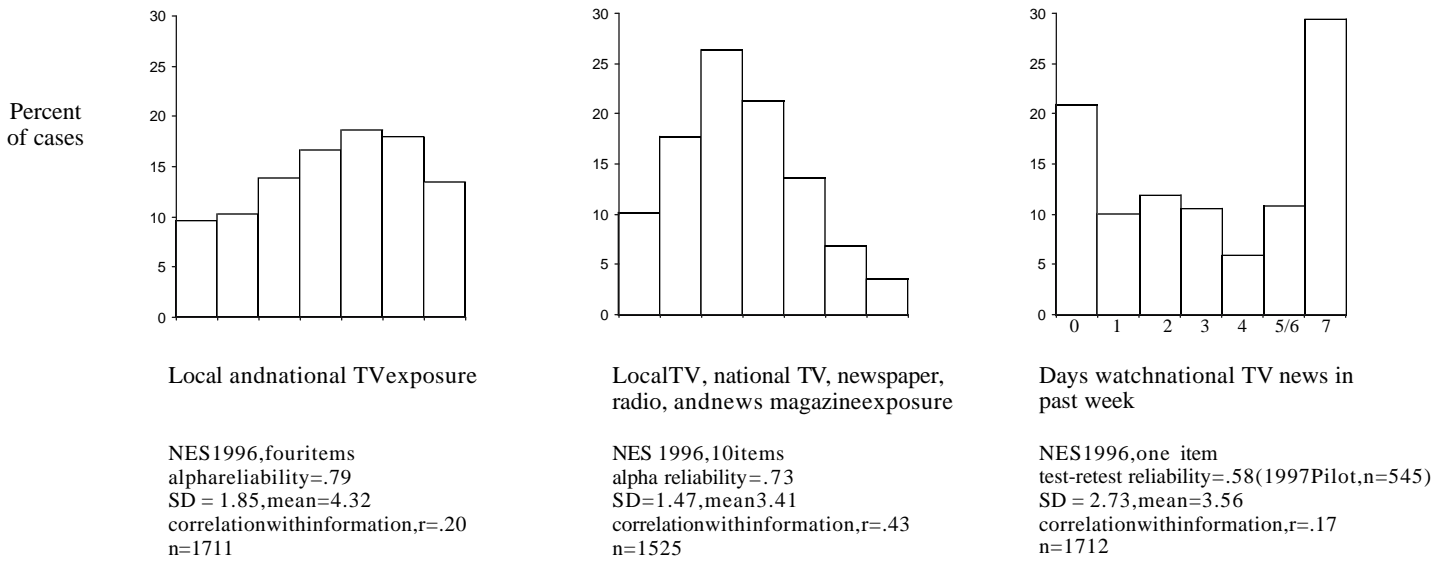
Figure 3 gives three examples of media scales that can be built from the 1996 NES study. In the left-hand panel is the distribution for a 4-item additive scale based on number of days watching local news, days watching national TV news, amount of attention to campaign stories on local TV, and amount of attention to campaign stories on national TV. These items make a coherent and focused scale with an alpha reliability of .79, but the scale correlates poorly with a 13-item measure of political information ( $r=.20$ ). Although some scholars disagree with me, I find it difficult to take seriously a measure of news exposure that has so weak an association with what ought to be the effect of news exposure.

INSERT FIGURE 3 ABOUT HERE

---

<sup>9</sup> I have argued elsewhere that political information is the best available measure of media exposure (Zaller, 1990; Price and Zaller, 1992). However, I do not see how simulation could shed any light on that issue and so ignore it in this paper. Information scales tend to have reliabilities equal to or lower than good media exposure scales, but to perform better in many kinds of tests.

Figure 3. *Three commonly used media exposure variables*



Note: The TV exposure scale is based on items v0242 to v0245. The omnibus measure adds items v0246, v0248, and v1333 to v1336. The scales were rebuilt from a principal components factor analysis and recoded to 7-point scales with intervals of equal width.

Source: 1996 NES study

A longer and more general measure of media exposure is shown in the middle panel of Figure 3. It includes two items each on national TV, local TV, newspapers, radio, and magazines, for a total of 10 items. All of the items from the previous scale are also in this one. The alpha reliability is lower (.73), but its correlation with information is somewhat better at  $r=.43$ .

These results suggest one problem in measuring media exposure. The reason the longer scale has a lower reliability is that few people attend to all media, preferring to specialize in just one or two. This leads to lower inter-item correlations and thereby lower alpha. The usual prescription in this circumstance is to use the shorter and more coherent scale. Yet, when a big campaign story occurs, it is generally covered in all media, and the analyst then wants to measure exposure to all media. A narrowly focused media scale does not achieve this purpose. Nor, typically, is it feasible to use a series of scales, one for each medium, since this mires the analysis in multicollinearity. These considerations favor the longer, omnibus scale. So also does the fact that the longer scale has better external validation in the form of a higher correlation with political information.

Another sort of difficulty is suggested by the right-hand panel of Figure 3, which shows how often Americans claim to watch the national TV news. As can be seen, about 30 percent claim to watch every day of the week. In a country of about 200 million adults, this translates into about 60 million viewers per night. Yet, according to Nielsen ratings, the combined daily audience for the networks plus CNN is around 20 million. And this is only part of the story. By also taking into account the people who claim to watch the news only one to six nights a week, we reach the startling conclusion that about half the respondents to NES surveys -- and by extension half the adult population of the country -- watch the national news on any given night.<sup>10</sup> This comes to about 100 million adults, or about five times as many as actually watch, according to Nielsen's more careful methods. This sort of overstatement is common in survey reports of media exposure. For example, a 1989 survey found that about 10 percent of respondents claimed to read the Wall Street Journal. This comes to 20 million readers, though the paper's circulation is around 2 million. Among the supposed readers, a few claimed to read it six or seven days a week, even though the paper publishes only five days a week.<sup>11</sup>

What might be the effect of such rampant overstatement on the ability of a measure of media exposure to detect campaign effects? A standard result of econometric theory is that systematic overstatement of scores on an X-variable does not bias coefficients (though it does bias the intercept). But systematic overstatement of X *in combination with* truncation of the high end of the X-variable is another matter. When this occurs, people whose true scores are low to middle on the X-variable can overstate their media exposure, but people whose true scores are at the top cannot. This leads to a

---

<sup>10</sup> Some respondents may have misunderstood the question, thinking that they saw national news on their local TV news program. A later question, however, permits correction for this misunderstanding. When respondents were asked which national news program they watched, some said they didn't watch any. Filtering out these respondents yields an estimate that 48 percent of respondents watch a national news program on any given night; without the filter, the estimate is 51 percent.

bunching up of dissimilar types in the top scale category, which fills up with a mix of middle-range people who are exaggerating and high end people who are unable to exaggerate. The expected effect is attenuation of the discriminating power of the scale at the high end. A probable example of such attenuation is shown in Figure 4.

INSERT FIGURE 4 ABOUT HERE

Figure 4 shows the percent of respondents who can identify the anchor of the TV network news programs they most often watch. For example, if a respondent said he watched the NBC Nightly News, he was asked which network Tom Brokaw worked for. People who said they never watched the network news, or were unsure which they watched, are excluded from the analysis.<sup>12</sup> Figure 4 shows how four measures of news exposure affect 'knowing your own anchor.'

By far the worst exposure measure is the item that ought to perform best, namely the one that asks how many days in a week the respondent watches the national news. This item is only weakly correlated with the ability to recognize the anchor of one's news show, and the people at the highest level of self-reported exposure are hardly more likely to get the right answer than people with the lowest exposure level. Thus, when TV news exposure and TV news exposure-squared are used in a logit model to predict anchor recognition, the term for exposure-squared is statistically significant at  $p = .003$ , two-tailed.<sup>13</sup> A plausible explanation for this odd pattern is that a high proportion of the people at the highest level of self-report are exaggerators and therefore fail to show the expected effect of high exposure.

Figure 4 also shows comparable results for three other measures. The broader measure of general news exposure has ten items, three of which are questions about "how many days in the past week" and probably suffer from some degree of over-report. This general measure also shows some non-monotonicity, though less than the TV exposure item alone.<sup>14</sup> Meanwhile, two measures that do not suffer from over-report bias -- education and an information scale based on objective tests -- show much stronger relationships with recognition of the name of one's news anchor.

Over-report bias, which no doubt varies across respondents and items, does not reduce calculated estimates of scale reliability. On the contrary, its tendency is to enhance apparent reliability. As long as survey respondents exaggerate with some degree of consistency from one exposure item to the next and one survey to the next, over-report bias cuts randomness and thereby enhances estimates of reliability. Over-report bias can thus hide the damage it does behind exaggerated reliability estimates.

---

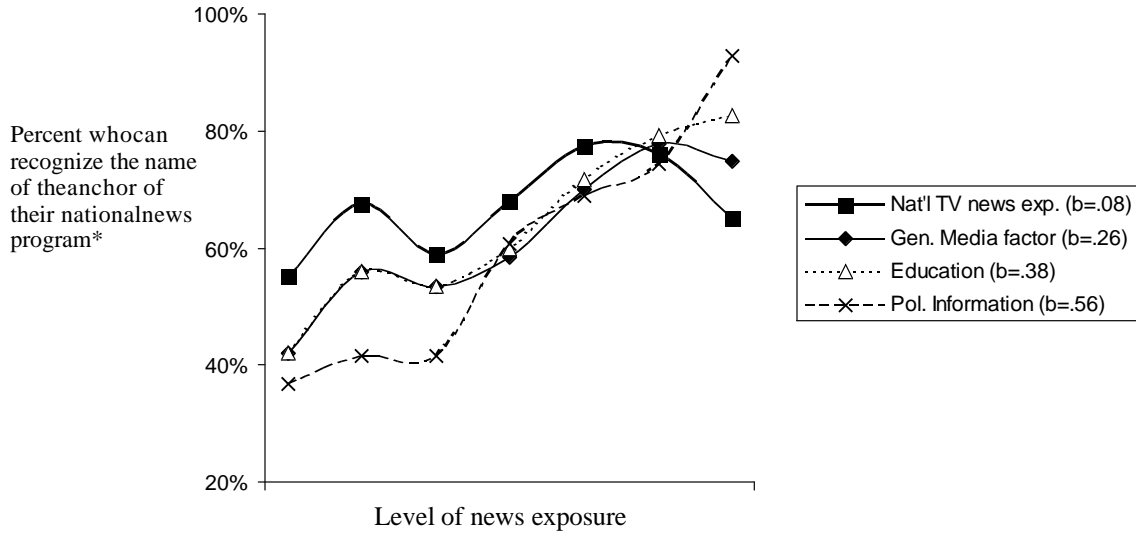
<sup>11</sup> Actually, 2 of 48, or 4.2 percent, but even so the anecdote is too good to pass up.

<sup>12</sup> Respondents were asked about the anchors for the three networks plus CNN, but I am analyzing only responses to the question about each person's own most watched program.

<sup>13</sup> When exposure and exposure-squared are used to predict anchor recognition, the term for anchor-squared is statistically significant at  $p = .008$ , two-tailed.

<sup>14</sup> In a logit like the one above, the exposure-squared term has a two-sided p-value of .08.

Figure 4. How four measures of news exposure affect learning the names of network anchors



\* We're interested in how much people learn about television news personalities. Take Tom Brokaw. Do you happen to know which network he works for-- is it CBS, CNN or which? ...Peter Jennings?...Dan Rather? ...Bernard Shaw?

**Note:** B's in key are unstandardized slopes from OLS regressions in which each independent variable has been scaled to a 1 to 7 range.

Source: 1996 National Election Study

From the NES data, then, there seem to be two kinds of media exposure scales: Broad or omnibus scales that have a low mean level of news exposure and exhibit no clear evidence of exaggeration, and more narrowly focused scales that have a higher mean and do exhibit evidence of exaggeration.

In the simulations that follow, I shall focus primarily on the broad type of exposure scale, for three reasons. First, it is what many and perhaps most researchers use when testing for exposure effects. Second, as I explained above, good reasons exist for preferring the broader kind of scale despite the apparently higher reliability statistics of more focused scales. Third, some survey respondents appear to exaggerate much more than others<sup>15</sup>; the simulation of a narrowly focused scale could accommodate this interpersonal difference, but it would require stronger assumptions than could be readily justified. I will later simulate an exaggerated exposure scale, but the exaggeration will be based on the simplistic assumption that everyone exaggerates to the same degree, and I will not rely heavily on this scale for any of the main conclusions of the paper.

Let me, then, begin building a media exposure scale. The first step is to assume a specific distribution of true media exposure scores. The one I have assumed is shown on the left of Figure 5. This distribution is not meant to recapitulate what gets measured in surveys; it is meant, rather, to depict what exposure scores would look like if they could somehow be observed independent of the random and systematic error that plagues survey measurements.

INSERT FIGURE 5 ABOUT HERE

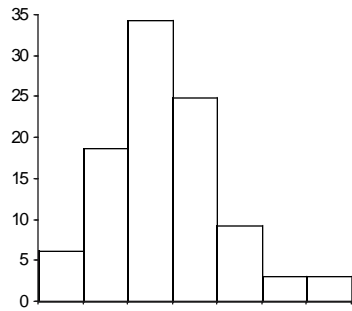
Although the true-score distribution I have assumed is partly guesswork, it is not unconstrained guesswork. Its most notable feature is a sparsely populated right tail. This reflects the assumption that I am simulating true scores on an omnibus scale -- a scale that covers everything from TV to newspapers to radio to magazines and that therefore gives respondents more opportunities to claim media use than any but a handful could honestly take.<sup>16</sup> In a scale intended to measure the gamut of media behaviors, the true score distribution should have a sparsely populated right tail. Moreover, I chose a true-score distribution that can, by the addition of random measurement error, be transformed into an "observed-score" distribution having properties similar to the NES omnibus exposure scale shown earlier. This observed-score distribution is shown as the middle distribution in Figure 3. As can be seen, the mean, standard deviation, reliability, and general shape of this distribution are similar to those of the NES omnibus exposure scale in Figure 1. The observed-score distribution in Figure 3 -- or, rather, the individual scores contributing to it -- will be my primary media exposure variable in the simulations that follow.

---

<sup>15</sup> As shown above, many people exaggerate their exposure to network news; yet, at the same time, Figure 1 also shows the existence of a large fraction of people who cannot be exaggerating (on the upside) because they say they never watch.

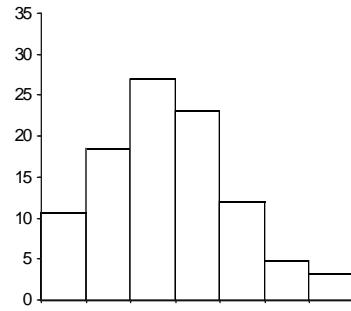
<sup>16</sup> Questions that might be used to create a scale with a comparably under-populated left tail -- e.g., "How many years has it been since you read a campaign story in a newspaper" -- are not very often asked.

Figure 5. *Simulated media exposure scales*



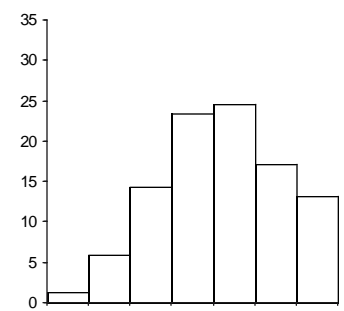
Assumed true-score distribution

SD = 1.32, mean = 3.34



Exposure scale with no exaggeration

True-score reliability = .70  
SD = 1.53, mean = 3.37



Exposure scale with exaggeration

True-score reliability = .66  
SD = 1.47, mean = 4.70

My method for creating this observed-score distribution was trial and error. Working on an excel spreadsheet, I created true-score distributions of various shapes, experimentally adding more or less randomness<sup>17</sup>, until I had a true score distribution that I could transform into an observed score distribution that closely resembled the NES observed score distribution.

The reader should be keenly aware that the properties of the two simulated distributions are consequential for all that follows. The true-score distribution specifies the X-variance that, as we saw earlier, affects statistical power. Thus, the assumption that true scores are bunched and skewed rather than spread out and symmetrical makes a difference. The amount of measurement error in the observed-score distribution also affects statistical power. If I chose a scale with more variance, or less measurement error, it would generate higher estimates of statistical power than the ones I report below. Hence, the reader is advised to ponder carefully the justification given for my assumptions in simulating true scores and observed scores.

Finally, the distribution in the right part of Figure 3 shows the scores that result when overreport and random measurement error are both added to the initial true-score distribution. The measurement error is the same as for the previous scale; the overreport is a constant 1.5 scale units for each person on a 7-point scale. I make limited use of this scale in the simulations that follow.<sup>18</sup>

Before going further, I must clarify my use of the term reliability, which is deployed in two different ways in the research literature. The most common usage is that of Carmines and Zeller (1979, p. 11-13), who say that "*reliability* concerns the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials." They distinguish reliability from validity, which is the extent to which a measure "measures what it purports to measure." And they add "just because a measure is quite reliable, this does not mean that it is also relatively valid." Thus, overreport of media exposure could, in their senses of the terms, make a media exposure variable more reliable but less valid.

A different conception of reliability occurs in Lord and Novick's (1968) classic treatment of *Statistical Theories of Mental Test Scores*. They write that "The reliability of a test is defined as the squared

---

Thus we are less likely to observe a sparsely populated left tail on an actual multi-item scale than a sparsely populated right one.

<sup>17</sup> Excel has a command for generating random numbers on a flat 0-1 interval. To create random measurement error, I added two of such random numbers, subtracted 1 (to center the sum on zero), and multiplied by 2.05. This yielded an error distribution with a range of -2.05 to +2.05, an SD = .84, and a roughly normal shape (except for fatter tails). A draw from this distribution was added to each individual's true media score in order to create an error-laden observed media score. The result was an observed score distribution having properties similar to NES observed scores, as explained.

<sup>18</sup> I note in passing that it is possible, through a careful choice of error structure, to transform the true-score distribution in Figure 3 into a scale having virtually the same mean, variance and reliability as the four-item NES TV exposure scale in Figure 1. But I concluded that, as indicated above, the assumptions necessary to do so are unavoidably arbitrary. Therefore, I confine my analysis to the simpler overreport scale discussed in the text.

correlation,  $r_{XT}^2$ , between observed score and true score (p. 61)." In this usage, the distinction between reliability and validity does not exist.

In this paper, I use reliability in the true-score sense of Lord and Novick rather than the repeated trials sense of Carmines and Zeller. Since my simulation generates both true scores and observed scores but no repeated measures, it is easy to use the r-square between them as the measure of reliability.<sup>19</sup> The true-score concept also seems more natural in cases, like my simulations, in which there is no validity concern. But the reader should note that the alpha reliability estimates calculated from NES data embody the repeated-trials approach and carry no implications whatsoever about validity. Thus if, as is likely, the NES data suffer from overreport bias and possibly other validity problems, an alpha reliability of .80 from actual data could be lower than a true-score reliability of .70 from simulated data in which there is no validity concern. Indeed, I assume that, if true-score reliabilities could be calculated for the NES data, they would be as low or lower than the true-score reliabilities I use in my simulations.<sup>20</sup>

In sum, I have simulated a primary media exposure scale that is based on a plausible distribution of underlying true scores, and that has observable characteristics quite similar to those of the media exposure scale many researchers would build from the 1996 NES survey. The most important of the parameters of the simulated scale are its true-score variance (as fixed by its distribution) and its true-score reliability.

There are now just two additional parameters of the simulation to fix -- the amount of residual variance in the vote equation we will use to test for media effects, and the amount of multicollinearity. Happily, neither presents much challenge.

#### Overall r-square of model

As explained earlier, the probability of obtaining a statistically significant coefficient for a given X-variable is higher, all else equal, if other X-variables are present in the model to hold down random error and drive up the overall r-square of the regression. For this reason, researchers are normally happy to load up their regression models with covariates, provided only that multicollinearity does not become an issue. In the context of presidential vote models, there may be ten, twenty, or even more covariates.

Yet, the last handful of variables typically explain little additional variance and may even reduce the adjusted r-square. The bulk of the heavy-lifting in presidential vote models is done by party identification, along with a few others. Table 1 gives results for a series of logit models used to predict the two-party vote in the 1996 NES study. For each of the 12 models for which summary statistics are reported in

---

<sup>19</sup> The simulation does not generate a measure based on repeated trials, but could readily be made to do so if there were reason.

<sup>20</sup> Alpha reliabilities are lower-bound estimates of reliability in the repeated-trials sense and would be higher if item difficulties vary, which they certainly do, especially in the general NES exposure scale in Figure 1. Even so, I suspect on the basis of much experience with the sort of data shown in Figure 2 that

Table 1, vote choice is the dependent variable, but the independent variables vary as indicated. For example, in model 1, party ID is the only independent variable; but in model 2, a set of standard independent variables is included, such as ideology and race, but party is not included. Each model involves a somewhat different configuration of the variables researchers might ordinarily use.

INSERT TABLE 1 ABOUT HERE

As the table shows, party identification alone can generate a pseudo r-square of .55 and correctly predict 88 percent of cases. When standard covariates such as issue positions and demographics are added, as in models 3 and 4, these figures rise to .66 and 91 percent. This level of fit for a dichotomous dependent variable ought to be high enough for anyone, but it is not unusual for analysts to add evaluations of the candidates on such traits as competence and compassion, and emotional reactions to the candidates on such dimensions as fear and pride. Occasionally one sees candidate feeling thermometers as independent variables. The addition of these variables to presidential vote models raises the pseudo r-square close to .80 and the percent of correctly predicted cases to about 95 percent. One may see higher values in some published articles.

Yet, as regards the detection of campaign effects, the models that achieve the highest r-squares come at a cost. The trait, emotion, and thermometer variables are so powerful, and so proximate to vote choice, that they absorb the impact of other variables. Thus, when the trait and emotion variables are added to the model that previously contained only party ID and the standard covariates, the r-square goes up but the coefficients for party and the other standard covariates go down. For the coefficients that are statistically significant in the initial model, the mean reduction in size is 36 percent. When thermometer scores are added along with traits and emotions, the overall reduction in size of previously significant coefficients is 46 percent.<sup>22</sup> If candidate evaluations were short-term causes of party identification, political ideology and other standard covariates, this would not be a problem; however, it is more likely that candidate evaluations are the effect rather than cause of these variables and hence absorb variance that “belongs” to them.

The same can no doubt happen in the case of campaign effects: If exposure to a debate or news story makes citizens more likely to vote for a candidate, it might positively affect trait judgments, emotional reactions, and thermometer scores, and these variables might then absorb the media effect. This danger would be especially great if media exposure were measured with more error than the other

---

the true-score reliabilities of the scales used in my simulations are higher than those of actual data. However, not all researchers may have this view.

<sup>22</sup> A useful point of reference: When party ID was added to a model that previously contained only the standard covariates, the average reduction of coefficients that had been statistically significant in the initial model was 18 percent. Reductions were greatest for ideology (50 percent) and cuts in government services (59 percent).

Table 1. Fit statistics for alternative models of the 1996 two-party presidential vote

Model:	<u>Pseudo r-square</u>	<u>Percent of cases correctly predicted</u>
1. Five-point party ID scale only	.55	88.3%
2. Standard covariates only*	.50	84.5%
3. Five-point Party ID Standard covariates	.66	91.0%
4. Dem. and Rep. Party ID dummies Standard covariates	.65	90.6%
5. Five-point party ID Standard covariates Traits battery	.76	93.9%
6. Five-point party ID Standard covariates Emotions battery	.75	93.8%
7. Five-point party ID Standard covariates Candidate thermometer difference	.79	94.2%
8. Five-point party ID Standard covariates Traits battery Emotions battery Candidate thermometer difference	.80	94.5%
9. Five-point party ID Traits battery Emotions battery Candidate thermometer difference	.78	94.8%
10. Five-point party ID 11. Traits battery Emotions battery	.75	93.9%
12. Five-point party ID Candidate thermometer difference	.77	94.0%

N of cases 1,034

**Source:** 1996 National Election Study

**Note:** The table reports summary statistics for each of 12 logit models. In each, major party vote in the 1996 is the dependent variable, but the independent variables are different across tests.

\* The standard covariates are self-described ideology; short scales on government spending and moral tolerance; one-item scales on national economic conditions, abortion, aid to blacks, government services, defense spending, job guarantees; respondent's race, age and gender.

variables, as it may well be. The problem can occur in any sort of model, but is especially likely to occur when summary variables, like trait evaluations or emotional reactions, are included in one's vote model.

In view of these considerations, analysis will focus on models having r-squares of .65.<sup>23</sup> Models having higher r-squares would show greater power to detect media effects, but since I cannot credibly simulate how candidate evaluation variables might absorb the effect of media exposure, the results would not be interpretable.

### Multicollinearity

If the variables used to predict vote choice in Table 1 are instead used to predict scores on the NES omnibus media exposure scale in Figure 3, the r-squares are about .14. The zero-order correlations between these media scales and vote choice are about .10. I shall therefore build relationships of these magnitudes into the simulated data.

The most important source of multicollinearity in the simulations, however, comes from use of interaction terms, as explained earlier. If, for example, I calculate the auxiliary r-square for the interaction term in Eq. 1 from simulated data, it is about .85. For the non-monotonic models, it is around .98. These are not values that I built into the simulation; they are values that have "fallen out" of the dataset as constructed to the specifications described above.<sup>24</sup>

## RESULTS

We have now developed specific empirical values for all of the general data characteristics that, as the earlier analysis indicated, affect statistical power in models of campaign effects. These are: the variance of true-score media exposure (as fixed by its assumed distribution) and the reliability of this scale; the residual variance in standard vote models, as measured by r-square; and the degree of multicollinearity between our key independent variable, media exposure, and other independent variables. We have also determined the size and shape of the particular campaign effects that may occur in U.S. presidential elections. We can therefore proceed with the simulation.

The essence of simulation is to create serial datasets in order to test how often particular relationships appear. And this is what I do. I create datasets of given sample size, that embody one of the campaign effects described above and that meet all of the data constraints described in the previous section. I then run a logit regression in each dataset in order to find out how often, over many such datasets, I am able to recover a statistically significant coefficient for the campaign effect I embedded in the data. I report two cutoffs for statistical significance -- .05 one-tailed and .10 one-tailed. The particular

---

<sup>23</sup> See appendix for discussion of the error term used to achieve this r-square.

<sup>24</sup> These r-squares are also comparable to what are obtained in actual NES data when a 10-item media exposure variable is interacted with an arbitrary 0-1 variable and made the dependent variable of an auxiliary regression in which media exposure, the 0-1 variable, and other "ideology" variables are used as independent variables.

models and coefficients of prime interest are those described above in the section on campaign effects. (For more details on how the simulations were done, see Appendix A.)

The simulations are structured so that half of the individuals have been interviewed prior to the campaign event and half afterwards, thus maximizing statistical power. Thus, if the N of a simulation is 2,000, it means there are 1,000 interviews in a baseline condition and 1,000 in the period following the media event. Note that the N's required to achieve a given level of power do not necessarily refer to the total size of the study; they refer to the cases available for a particular test. Typically, the available N will be smaller, perhaps greatly smaller, than total N. Thus if, for example, one were testing the impact of exposure to a presidential debate in an NES survey of 2,000 respondents, one probably could not use all 2,000 cases in the test. Rather, one would probably compare opinion in the week or so before the debate with opinion in the week afterwards. Given the NES design, this might involve, at most, 500 respondents from the total sample of 2,000.

The results of the simulations are shown in Figure 6. The panels are arranged so as to correspond to the layout of Figure 2, which showed the shape and size of the campaign effects under study.

INSERT FIGURE 6 ABOUT HERE

The most general finding in Figure 6 is that detection of exposure effects is likely to be unreliable unless the effects are both large and captured in a large survey. Surveys, or subsets of surveys, having fewer than about 2,000 cases may be unreliable for detecting almost any sort of likely exposure effect, and even surveys of 3,000 could easily fail to detect politically important effects.

Let me take an example. The 2000 Democratic convention produced a 7.3 percent bounce for Al Gore.<sup>25</sup> In political terms, this was huge shift in vote preference. But if researchers had interviewed 1,000 people in independent samples before and after the convention, for a total of 2,000 respondents, and if the bounce was non-monotonic with respect to media exposure, the chance of detecting it at the .05 level, one-tailed, would be about 45 percent.<sup>26</sup> And this assumes a media exposure scale comparable in reliability to the 10-item scale carried in the 1996 National Election Study.

But this may be overly blunt and pessimistic. A more differentiated summary of the findings in Figure 6 may be offered as follows:

- Some large campaign effects -- linear exposure effects of about 10 percent, which is the size of a large convention bounce -- can be reliably captured in election surveys of 2,000 respondents. Elite

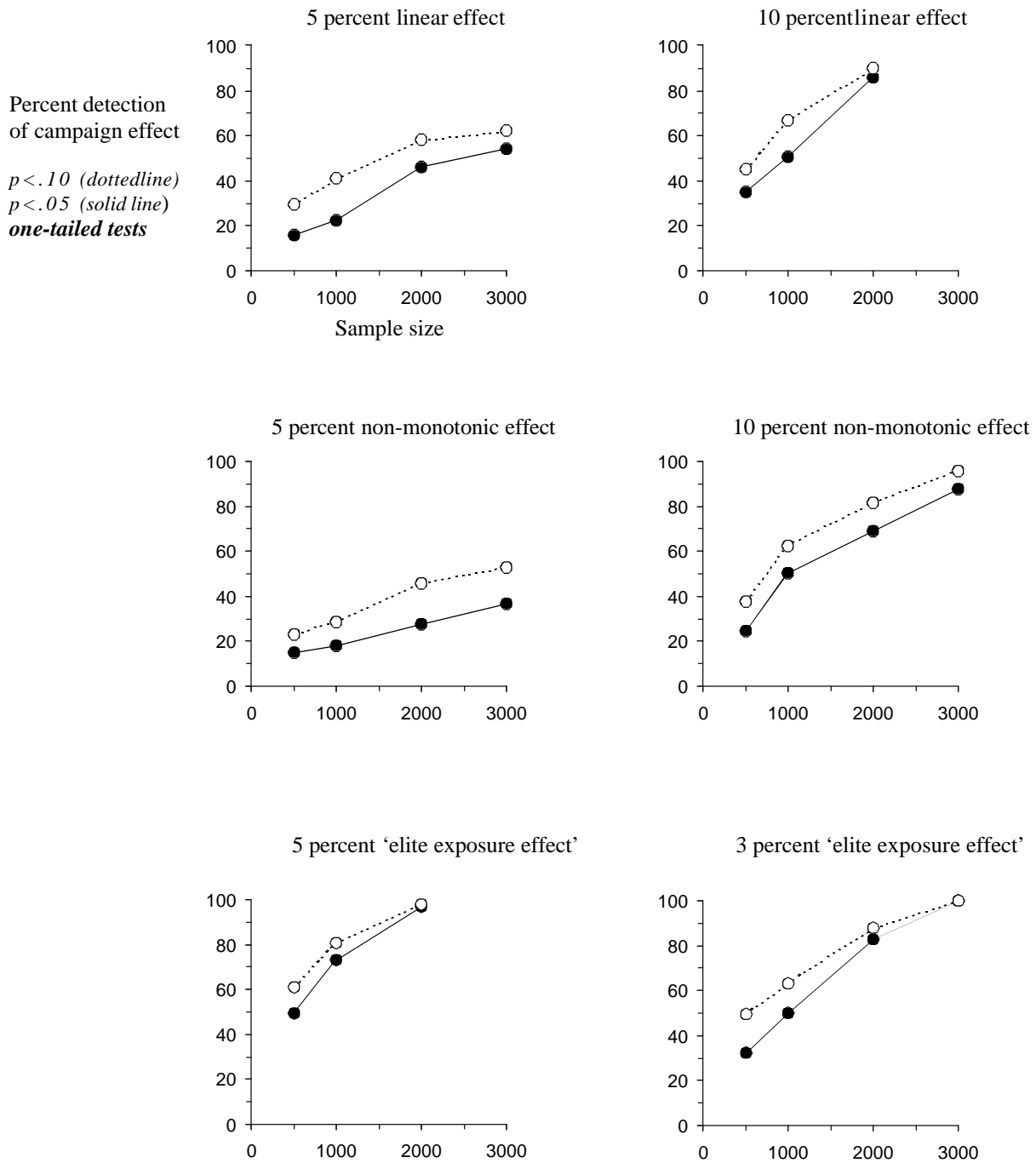
---

<sup>25</sup> This is the average Gore gain in six polls reported in "In Postconvention Polls, Gore Bounces Just Ahead of Bush," Adam Clymer, *New York Times*, August 22, 2000, p. A18.

<sup>26</sup> This result has been interpolated from the 5 percent and 10 percent non-monotonic subpanels of Figure 6 below.

<sup>28</sup> I did not, incidentally, expect the elite exposure effect to be so detectable when I chose it for study. I chose it because my sense, from various data I have examined, was that it was relatively common, as in Figure 1. However I now suspect that I had this sense because this kind of effect is so easy to capture in surveys. Five percent non-monotonic effects could occur just as often and remain invisible.

Figure 6. Detection rates for campaign effects of given size and shape



**Note:** Size and shape of campaign effects are as given in Figure 2 and accompanying text. Each point in each graph is based on at least 50 trials at given sample size, and the vast majority are based on 100 to 200 trials.

exposure effects of 5 percent can also be reliably captured with 2,000 respondents, and can be fairly reliably captured with only 1,000 respondents.

- Linear and non-monotonic campaign effects of normal size -- a five point shift, which is equivalent to a gaffe or large debate bounce -- cannot be reliably captured in any reasonably sized election survey. A separate series of runs, not shown in Figure 6, shows that with 5,000 respondents, a 5 percent non-monotonic effect can be detected at the .05 level with only about 50 percent reliability.
- The big surprise of the simulations is the relatively high level of detectability of what I have called "the elite exposure effect." A three percent elite exposure effect is more detectable than a five percent linear or non-monotonic effect, and nearly as detectable as a 10 percent linear effect. (The reader should recall that linear exposure effects, as I have tried to realistically define them, include a "discussion effect" that is independent of individual media exposure.)<sup>28</sup>

One of the features of these simulations is that I have used a particular model to create an effect and then tried to recover the effect with that same model. But the question arises: what if, as is essentially always the case, scholars are unsure what the right model is -- that is, what model has generated the campaign effect they are trying to detect? And what if they then use a "wrong" model instead -- a model that seems plausible but is not the one actually at work?

Because one always knows the "right" model when doing simulations, it is possible to address this question.

Let me begin with the non-monotonic exposure model. As explained earlier, I used a slightly "wrong" model in my attempts to recover the non-monotonic exposure effect. I created the effect with a function that uses media exposure and media exposure-cubed, but attempted to recover it with a model using media exposure and media exposure-squared. I did this because I felt that, although scholars might perhaps try a model with media exposure-cubed in it, they would be reluctant to report an x-cubed effect for fear that it would raise the specter of ad hoc data mining and overfitting. It would be better to write about something else than to use an x-cubed term simply because it worked when nothing else would. (I, personally, would feel this way.)

But how serious is this problem? In the case of the non-monotonic exposure model, it is not serious at all. In a set of 100 trials of sample  $n=1,000$  and a 10 percent effect, a model with a media-squared term had this term significant at the .05 level 44 times. For the same 100 samples, a model with a media-cubed term had it significant 45 times.

I made a similar test with the cubed version of the elite exposure model. Here the results suggest a possible advantage to the unconventional models. For 100 trials on samples of 2,000, the correct media-cubed interaction term captured the campaign effect 76 percent of the time, as compared to 77 percent for a media-squared interaction term and 66 percent for a linear media interaction term.

More simulations are needed to pin down this result. But on the assumption that it holds, the question arises: Why does having the "right" model not make more difference in these cases? The answer, I suspect, is measurement error on the media exposure variable. Degradation of the independent variable takes the edge off – flattens -- its relationship with the dependent variable, so that a "wrong" linear or square term works almost as well as the "right" cubed term. This, however, is only speculation.

Given the low power of most surveys to detect non-monotonic effects when they occur, a related question arises: If researchers test the "right" non-monotonic model but fail to find support for it, what do they do next? The answer, in many cases, is that they revert to a less complicated model.

The results of doing so may often give a misleading impression. When a five percent non-monotonic effect has occurred, the chances of getting support for equation 2 are only 18 percent for a sample of 500 and 28 percent for a sample of 1000. But a model that drops the media exposure variables and uses Time as the sole campaign variable gets significant coefficients for Time with 62 and 88 percent frequency, respectively. This is despite the fact that Time ought to have no effect apart from its interaction with media exposure (see Figure 2).

Researchers who fail to find support for the model they expected to hold, but do find support for a simpler and perhaps less likely model, will normally report results for the model that "worked," perhaps mentioning in passing that the expected model failed to work. What else can they really do?

Actually, there is something. Researchers who obtain null results for their expected model could and, I suggest, probably should attempt to assess the statistical power of their test. "What," they should ask, "are my chances of finding statistically significant support for my expected model, given the amount and reliability of data that I have to work with?" I return to this matter in my concluding remarks.

One final set of results. I noted earlier that many media exposure scales are afflicted by exaggerated statements of usage. I tested the effect of such over-report by simulating a media exposure scale with the same true score and random error variance as my primary scale -- but with a 1.5 point across-the-broad over-report added to true scores on a seven-point scale. Persons who scored at the top of the true exposure scale without the over-report could not go over the top, but were held down by a ceiling effect to the top bracket.

On the basis of test runs, it appeared that over-report did not undermine power to detect either linear effects (where power was low to begin with) or elite exposure effects (where tests were not very sensitive to form of model either). However, it did appear that over-report had an important effect on tests of the non-monotonic model. For a sample of 1000 and an effect size of 10 percent, power was reduced from 50 percent with the primary media scale to 27 percent with the scale having over-report. When the

sample size was raised to 3,000, the effect of over-report was to reduce power from 88 to 51 percent. These tests were based on 100 trials per cell.

#### CONCLUDING COMMENTS

In opening a 1993 article in the American Political Science Review, Larry Bartels wrote:

The state of research on media effects is one of the most notable embarrassments of modern social science. The pervasiveness of the mass media and their virtual monopoly over the presentation of many kinds of information must suggest to reasonable observers that what these media say and how they say it has enormous social and political consequences. Nevertheless, the scholarly literature has been much better at refuting, qualifying, and circumscribing the thesis of media impact than at supporting it. (267)

Bartels' attributes this state of affairs in part to scholarly "carelessness" about the consequences of measurement error and in part to "limitations of research design." The limitation I have sought to bring into sharp relief in this paper is the size of surveys. When the data available to test models of media effects have three times more power for supporting "wrong" models that have no media variables than for "right" models that do, it cannot be good for communication studies. One must wonder, in particular, how many of the null findings that pervade the study of mass communication are actually cases in which the truth (if it could be known) is a five percent non-monotonic effect that cannot be detected in any reasonably sized sample and so registers in the data as a null effect. It is hard even to make a plausible guess about how often this might happen.

What, then, is to be done?

The easy answer is that funding agencies should give scholars more money to do bigger studies. With the kinds of analyses reported in this paper, scholars can write grant proposals that make a more persuasive case for big surveys than has been made in the past, and funding agencies will respond by opening their checkbooks.

Although I think scholars who design surveys should devote more attention to power than they typically do -- what is now typical is close to none -- I doubt that simply dropping a power analysis or two into a grant proposal will have a dramatic impact on support. For one thing, the need for cases must always be judged in light of particular purposes. If, as suggested by my analysis, some important communication effects cannot be reliably detected even with 5,000 cases, funding agencies might insist that scholars give up the Holy Grail of explaining the effect of media exposure on vote choice and focus instead on more realistic goals. Moreover, no case for bigger surveys is likely to be persuasive if it is confined to grant proposals and methods papers like this one. It will be persuasive only when a large fraction of the scholarly community internalizes the lessons of methodological critiques concerning power, measurement error, modeling sophistication, and other matters, and demonstrates by the collective

weight of what it publishes in the top journals that we really are at the limits of we can accomplish with studies of conventional size. We are a long way from this situation.

It would also seem that the scholars who consume surveys should make more use of power analysis in their studies, especially in studies in which they turn up null or trace effects. That is, they should estimate the probability of rejecting the null hypothesis with respect to the particular size of effect they expect, given the model they are using, the characteristics of the variables in it, and the amount of data they have. The first to profit from such analysis would be the scholars themselves, who will presumably find it easier to publish null findings if they can show that their analysis has been sufficiently powerful to detect a real effect if one were present. Readers of their research would obviously also appreciate this sort of information.

Most scholars seem, like other humans, to be intuitive believers in what Kahneman and Tversky have called "the law of small numbers": If an effect is real, it will show up even in relatively small data sets. Such scholars might profit from pondering Figure 7. In repeated attempts to recover a 10 percent non-monotonic effect from samples of 1,000, I succeeded at the .01 (one-tailed) level 22 percent of the time and at the .05 level for an additional 22 percent of the time. Yet I also failed to make a weak .25 (one-tailed) cutoff 16 percent of the time and failed to make even a .40 cutoff 14 percent of the time. Further, in the 50 trials on which these figures are based, I got sign reversals on the critical "time X media squared" term 5 times, or 10 percent of cases.

INSERT FIGURE 7 ABOUT HERE

As we have seen, 10 percent is a big campaign effect and 1,000 cases is a reasonably large sample size. Researchers often deal with smaller effects and smaller samples or subsamples. Thus, it would be quite interesting to know how many of the studies that have fed the "minimal effects" orthodoxy of the past five decades have had the power to detect medium or large effects if such effects had been present.

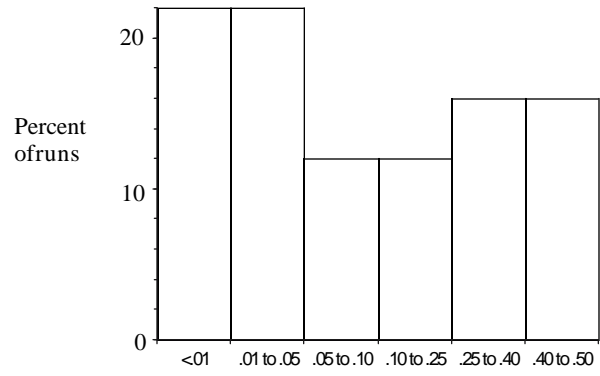
As important as it is, the incorporation of power analysis into substantive research reports cannot be expected to happen quickly. Indeed, if the researchers must go through as much work as was necessary to produce this paper, it may take a very long time. A priority for future studies of mass communications should therefore be to figure out how to accomplish power analysis more efficiently.

For certain purposes in communication studies, and in political science more generally, the formulae set out in Jacob Cohen's Statistical Power Analysis in the Behavioral Sciences (1988) would give accurate forecasts of power. These formulae and more have been incorporated into a highly accessible computer program, PASS2000 (NCSC, 2000), which, I believe, should be used much more frequently in behavioral research than it is. There is, however, a significant impediment to routine use of power formulae in communication studies and elsewhere: Figuring out expected effect sizes and translating them into increments of r-square, as required for input into power formulae for regression models.

---

<sup>29</sup> This is my paraphrase.

Figure 7. *Distribution of one-tailed p-values for tests of 10 percent non-monotonic effect in samples of 1,000*



N = 50 trials

Indeed, it would be hard to get such values without going through most of the steps I have gone through in this paper.<sup>30</sup> Suppose, for example, that my Equation 1 were a simple OLS model, for which analytical calculations can be made. How would a researcher specify, a priori, the expected increment in r-square (net of multicollinearity) of the  $b_3$  coefficient, so that the power of a given dataset to capture it could be calculated? Only by simulation, I suspect. And simulation is, as this paper shows, a rather cumbersome undertaking.

But even if effect size were easy to specify and power calculations extremely easy to carry out, scholars might still be slow to adopt power analysis. Indeed, it is a virtual certainty they would be. I say this because of experience in other scientific domains in which power analysis is, in fact, much easier to do and yet still has made only limited headway. For example, it is quite easy to use power analysis to determine how many subjects are necessary to observe a mean difference of given size in an experimental study with a dichotomous treatment variable. Yet in experimental studies from medicine to psychology, such calculations were slow to become routine. For example, a study in the 1994 Journal of the American Medical Association calculated the power of 102 randomized experimental trials of new drugs or treatments that had turned up null results. It turned out that only 16 percent of these studies, all of which had been published in top journals, had the power to reliably detect a 25 percent experimental effect on their dependent variable, and only 36 percent could reliably detect a 50 percent experimental effect. (Does anyone want to guess what the average power of studies that have found null communication effects might be?)

There are, to be sure, signs of progress. Yet, after reviewing them in the preface to the second edition of his power analysis textbook, Cohen commented:

It is clear that power analysis has not had the impact on behavioral research that I (and other right-thinking methodologists) had expected [when he published the first textbook on power analysis in 1969]. But we are convinced that it is just a matter of time. (1988, p. xiv)

Those of us who have designed studies intended to measure communication effects, as well as those of us who have used the data generated by these studies, have been among the laggards in taking power issues seriously. It would behoove us, as workers in a subfield full of minimal and null effects, to get to the vanguard.

---

<sup>30</sup> The actual inputs to power formulae are r-square with and without the test variable(s) in the model. No direct value for multicollinearity is used; however, the difference in r-square is strongly affected by the degree of multicollinearity between the test variable and other X variables in the model. Hence, multicollinearity must be taken into account when calculating the difference in r-square. When multicollinearity is high, it may be hard to get a large increment in the r-square even when the effect of the test variable on Y is large.

## APPENDIX A

The simulation of data was done as follows: Each of several thousand "individuals" on a spreadsheet is assigned a fixed value on a "true media exposure scale." The values on this scale run from 1 to 7 and have the distribution shown in Figure 3. Each individual is also assigned a "measurement error" value, which, when added to her true media score, generates a new "observed media" score. Each individual is also given a Time score of 0 or 1 to indicate whether she was interviewed before or after the media event. So far, then, we have four columns of variables -- true media exposure, time, measurement error, and observed media exposure (as sum of two previous columns).

Next, each individual is assigned two "ideology" scores, each of which is a separate variable. These scores are based on random numbers on a uniform 0-1 interval. To the first ideology score is added a fraction of each person's observed media score, where the fraction is chosen by trial-and-error to induce the necessary amount of multicollinearity with observed media exposure in the final vote equation. The two ideology scores are then combined, along with an additional dose of random error, to create a "presidential preference" scale. The amount of this random error is chosen to produce the desired overall r-square for the vote equation, .65. The form of the random error is a uniform distribution.<sup>31</sup>

The preference variable is a continuous scale meant to simulate individuals' underlying preferences for the two candidates. It runs from about -2 to +2, with a mean of 0 and an SD of .8. Thus, three more columns of variables are added to the spreadsheet.

The preference scale is ultimately converted to a 0-1 vote variable and then discarded. In view of this, I see no reason in statistical theory that the underlying presidential preference scale must match aspects of real data. Yet prudence suggests that I make sure its distribution does not stray too far from actual data. For this reason, Figure A has been created to compare its distribution with that of the feeling thermometer difference scores of the two candidates in the 1996 election. Despite the pro-Clinton skew of the 1996 data, the shapes of the two distributions are roughly similar. When both scales are coded to a common 11-point range, their SDs are reasonably similar, as the figure shows.

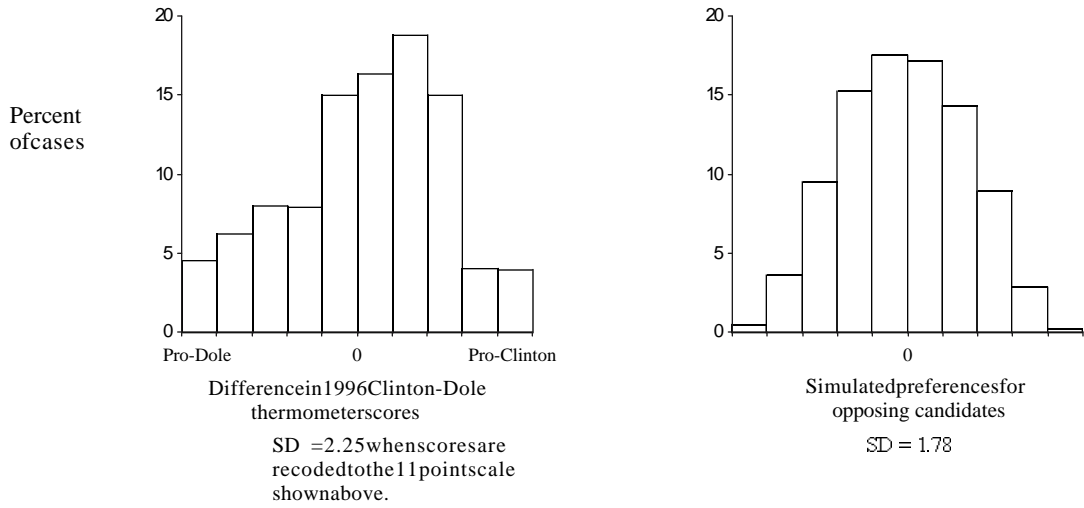
INSERT FIGURE A ABOUT HERE

Before the preference scale is recoded and discarded, it is "shocked" to create a campaign effect. For individuals who are 0 on Time, there is no shock. But everyone who scores 1 on Time has her continuous preference score adjusted to be somewhat more favorable to Candidate A. The amount of the adjustment is in relation to true media exposure scores, but the particulars of the adjustment depend

---

<sup>31</sup> The conventional assumption would be normal errors. However, normal errors are at least as unlikely as uniform errors. I did a sensitivity test with errors formed as the sum of two uniform distributions, which produces a very normal-looking (though not quite actually normal) distribution. It turned out that it made

Figure A. Actual and assumed distribution of voter preferences



Source: 1996NES

on shape and size of the media impact to be created. For example, if the aim is to create a non-monotonic campaign effect, people with middle range true media exposure receive the biggest shocks.

Note the use of true media scores to create the shocks: I adjust candidate preferences on the basis of true media exposure, but I later attempt to recover the media effect with error-laden observed media scores. This is an obviously realistic feature of the simulation.

For candidates who are extremely pro- or anti-Candidate A to begin with, the shock makes no difference, because they do not cross the threshold between the two candidates. But for persons whose preferences are initially just below the threshold of support for Candidate A, the shock may make a difference. The size of the shock is chosen to create the desired amount of aggregate change -- a three, five or ten percentage shift toward Candidate A.

Finally, the candidate preference scale is converted to a 0-1 variable indicating a vote for Candidate A or his opponent. This vote variable is then used as the dependent variable in the analysis.

New randomness terms are added at several points in each new dataset: The assignment of scores on the two "ideology" variables, the assignment of measurement error, and the assignment of a random term to the continuous candidate preference scale. In each of the many artificial datasets I create, individuals' observed scores change but basic relationships in the data remain, in expectation, the same.<sup>32</sup>

The Excel spreadsheet used to create the 10 percent linear effect is available on my personal webpage as "powersheet" at [www.sscnet.ucla.edu/polisci/faculty/zaller/](http://www.sscnet.ucla.edu/polisci/faculty/zaller/).

#### ACCURACY OF SIMULATIONS

To check of the accuracy of my general approach, I did power simulations for cases for which power formula could also be used. That is, I did the regression equivalent of 1) simulating the SE of a survey of  $p=.4$  and  $N=1000$  through computer-generated coin-tosses and 2) comparing the result to the SE as calculated by the standard formula. The simulations were in general accord with the results obtained from the standard regression formula. I also confirmed that my statistical program, STATA, produced standard errors for coefficients that were generally consistent with the SD of coefficients obtained from repeated trials. By generally consistent, I mean that sets of simulations were within five to 10 percent of expected values, and that results did not seem, on visual inspection, to be consistently above or below expectation.

---

no difference: For a 10 percent linear effect on samples of 1,000, detection at the .05 level over 250 trials was 48 percent for uniform errors and 50.8 percent for the second error structure.

<sup>32</sup> In a classical Monte Carlo experiment, the ideology scores also would have been fixed from trial to trial rather than, as here, drawn fresh in every simulation. I doubt this feature will make any difference, and it is somewhat cumbersome to implement in my setup. The media exposure variable is fixed X, except for measurement error, which is freshly drawn for each simulation.

## References

- Bartels, Larry 1993. Messages received: the political impact of media exposure. *American Political Science Review*, 87,267-286.
- Carmines and Zeller 1979 *Reliability Analysis*. Sage. 1979.
- Cohen Jacob 1988. *Statistical Power Analysis in the Behavioral Sciences*.
- Converse, Philip 1962. Information flow and the stability of partisan attitudes. *Public Opinion Quarterly*, 26, 578-599
- Holbrook, Thomas 1997. *Do Campaigns Matter?* Sage.
- Hsieh, Block, and Larsen (1998) Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. "A Simple Method of Sample Size Calculation for Linear and Logistic Regression," *Statistics in Medicine*, 17, 1623-1634.
- Hunt, Mark and John Zaller 1995. The Rise and Fall of Candidate Perot: The Outsider vs. the System, *Political Communication*, 12.1: 97-123.
- Johnston, Richard, Andre Blais, Henry Brady and Crete 1992. *Letting the People Decide*. Stanford.
- King, Gary, Robert Keohane, and Sidney Verba, 1997. *Designing Social Inquiry*. Princeton University Press.
- Lord, Frederic and Melvin Novick's 1968. *Statistical Theories of Mental Test Scores*, Addison Wesley.
- Moher, D. and G.A. Wells 1994. Statistical Power, Sample-Size, and Their Reporting in Randomized Controlled Trials," *Journal of the American Medical Association*, 272: (2) 122-124.
- Price, Vincent and John Zaller 1993. Who Gets the News: Measuring Individual Differences in Likelihood of News Reception, with Vincent Price, *Public Opinion Quarterly*, 57, 133-64.
- Shaw, Daron 1999. A study of presidential campaign event effects from 1952 to 1992. *Journal of Politics* 61, 38-410.
- Zaller, John 1990. "Political Awareness, Elite Opinion Leadership, and the Mass Survey Response," *Social Cognition*, 8, 125-153
- Zaller, John 1992. *Nature and Origins of Mass Opinion*. Cambridge.