TESTING TELEVISION ADVERTISING USING INTERACTIVE TELEVISION

The effectiveness of political advertisements

Josh Clinton
John S. Lapinski

This paper introduces several new methods are aimed at improving ad testing. The power of interactive television and scientific statistical techniques are leveraged to create a unified methodology to test ads. The methodology makes three specific improvements, including the utilization of a random probability sample, the development of a system that allows for self-administered ad experiments given in a respondent's own home, and the employment of a fully randomized, panel experimental research design. The paper draws on empirical evidence from the Yale Advertising Study, focusing on measuring the effect of political ads on voter attitudes and behavior. The study included 12,350 interviews.

INTRODUCTION

This paper introduces several new methods that explicitly measure the effectiveness of television ads – where an ad's effectiveness is measured by its ability to produce or facilitate the desired behavioral outcome in its viewers. The importance of determining the effectiveness of an ad is perhaps best expressed in David Ogilvy's "18 Miracles of Research" which warns its readers that "Advertising people who ignore research are as dangerous as generals who ignore decodes of enemy signals" (Ogilvy 1985, 158). It is only through well thought out, scientific research that we can assess the true effectiveness of an ad, and consequently, determine which ads should be used.

Given the amount of money spent on advertising, combined with increasing competitive market pressures, it is more important than ever for executives to ensure that money spent on advertising is spent on *effective* advertising. Nearly $50.5 billion was spent in 1999 on broadcast and cable television advertising in the United States alone.[1)] Add the advertising dollars spent in other mediums, and it is clear that the effective testing of ads is critical both financially and to increase brand share.

In this paper, we present a new methodology for testing the effectiveness of ads that uses the technology advancements offered by interactive television. Although the methodology is applied to a political context, every aspect is directly applicable to a commercial context. The scope of the methodology we present, and the technology that we utilize, is expansive, as it greatly improves upon our current methods of assessing the effectiveness of any product promotions conducted through a video or audio medium (e.g., TV, web or radio). In fact, our methodology generalizes even further, as it is easily adapted to assess complete television shows or movies. In sum, the methodology we present can be used to test any type of multi-media based ad.

Our methodology differs in several respects from other attempts that seek to measure an ad's effect on product sales. For example, in contrast to a "single source" research approach (Jones 1995; McDonald 2000; Helgesen and Micalsen 2000; Battais and Spitzer 2000), which measures an ad's effectiveness by examining the amount of sales associated with an ad *after* it is aired, our procedures allow us to determine the effectiveness of an ad *before* it is shown commercially. Given the rising costs of advertising, the ability to assess an ad's effectiveness during a pre-test phase has the possibility of producing significant cost-savings.

NEW INNOVATIONS

Our methodology offers three methodological improvements over standard ad testing techniques (e.g., "dial group" research). Current techniques, even those based on experimental methodology, suffer from tremendous problems of both internal, and more importantly, external validity.[2)] We capitalize on recent developments in interactive television to present a methodology with both internal and external validity.

We provide three design improvements that unilaterally increase a researcher's ability to *scientifically* determine the effectiveness of an ad. First, although we administer our surveys via the internet, our experiments utilize a random probability sample of the United States adult population.[3)] This is a great improvement over using "convenience" samples. Second, the setting for our experiment is the most natural possible. Most studies, especially those experimental in nature, conduct research either in sterile classrooms, offices or modified rooms designed specifically to resemble family rooms or kitchens. This is attributable to the difficulty associated with showing multimedia ads in the context of an experiment. In contrast, our self-administered experiments are conducted in the respondent's own home. Furthermore, the ads are viewed on the respondent's own television, creating a natural setting for watching commercials. Because of these two advances, our research is the *only* study able to test the effect of ads on a random population sample in a natural environment.

The final design improvement is the experimental panel design that we utilize. Our panel design consists of a pre-treatment survey, experimental survey, and post-treatment follow-up survey. This allows us to determine not only how effective ads are initially, but perhaps more importantly, it allows us to measure the "staying power" of ads. In effect, we are able to track changes attributable to the ads on an individual – not aggregate – level across time. We now examine these three improvements in greater detail, emphasizing the benefits associated with each.

Methodological Improvement 1: Random Probability Samples

The majority of current ad testing relies on convenience samples. For example, most academic studies rely on undergraduate students willing to participate in studies for cash payments and/or school credit. Although researchers write extensive justifications for why their convenience samples possess better external validity than might be expected, convincing evidence is never supplied. Often researchers try to increase the external validity of the study by supplementing focus groups with individuals recruited from other places.

However, physical location still plays a critical role in recruiting, as drawing respondents from "other places" is not a substitute for random selection. For example, major political/market consulting companies such as Greenberg Research relied on small groups (less than 50) in a few select "battleground" locations to dial test ads in support of Democratic presidential candidate Al Gore. Although a considerable amount of effort was put into making sure that a demographically diverse group of people was included in the groups, participants were not randomly selected.

The importance of a random probability sample cannot be understated, as without a random probability sample, we know very little about the external validity of results. In other words, without a random probability sample we rely on intuition instead of a scientific foundation. Random probability samples allow us to obtain an estimate of the opinions of an entire population of interest (e.g., the entire adult population, all registered voters, all male primary household shoppers between the ages of 35 and 42) without having to interview everyone in the population. This is a powerful tool, which works as follows: A small sample (i.e., subset) of the population of interest is selected to estimate the unknown opinion of the entire population of interest. The unverifiable *claim* (absent a census of the population of interest) that the researcher must make is that the opinion of the sample examined is an unbiased estimate of the population's opinion.

Random sampling becomes important when this claim is made. Random sampling – whereby every member of the population of interest has an equal chance of being included in the sample – is the device that enables us to mathematically prove the claim that the sample opinion is an unbiased (and therefore accurate) estimate of the population's opinion (Kish, 1995; Lohr, 1998). Without random selection, it is *impossible* to determine the generalizability of the results. In other words, without random selection, there exists *no scientific basis* for extrapolating the results from the sample to the population of interest. The bottom line is that without random probability sampling it is impossible to determine the "margin of error" or accuracy of the testing.

The importance of random sampling should not be underestimated, even though it often is unfortunately either ignored or impossible within some types of market research (e.g., mall-intercept studies, internet surveys). In the absence of random sampling, the results summarize *only* the opinions of those particular respondents. This is true regardless of how demographically similar the sample is to the population.

While random sampling is a necessary condition for scientific validity, until recently its cost has been prohibitive. In addition, it was difficult, if not impossible, to randomly administer ad tests to geographically dispersed

populations (e.g., nationwide tests). As briefly summarized above, the "work-around" procedure would be the creation of special facilities where people from the surrounding areas could be recruited to go to the facility to view and assess the ads in question (e.g., "mall-intercept" surveys). Putting aside issues of non-response, it is clear that any sample resulting from this procedure is non-random, if only because people living far from the facility are ineligible.

Methodological Improvement 2: Self-Administered Ad Experiments

A lab or classroom converted into a pseudo-living room with television is clearly a different environment than one's own home. By conducting research in such a specially designed environment, even absent the problems of convenience sampling discussed above, the researcher introduces bias into the results. However, it is impossible to know the magnitude and direction of the biases. Researchers usually dismiss the problem by ignoring it, largely because a solution has not existed until now. To address the problem, and thereby minimize any resulting "interview" bias, we present the ads in the context most akin to that context in which it would be experienced normally.

We use an interactive television device set up by the respondent in the respondent's own home. This allows us to test the effects of ads in arguably the most natural consumer environment. Specifically, our surveys are administered over the internet to panelists using Microsoft's WebTV interactive television. The interactive television device enables respondents to take surveys in the comfort of their own home on their own television set at a time convenient to them. An added benefit is that respondents do not have to wait for videos to "stream" or download. Instead, WebTV's technology allows us to automatically download ads onto the WebTV's hard-drive before the survey is administered. This downloading is done automatically, and without any required action on the part of the respondent.

Since the ads reside on the hard-drive of the respondent's WebTV device, they can be shown, on-demand, instantly. The delivered ads are high-resolution, and of television quality, making it virtually impossible to distinguish between a regular television ad and what is shown in our experiments. Insofar as taking the survey in a more natural environment encourages respondents to provide opinions unbiased by the events associated with the survey (e.g., the interview), our experimental surveys' results are more accurate than those conducted in labs.

Methodological Improvement 3: Experimental Panel Design

Experimental designs are powerful because they enable researchers to establish causal arguments. For example, by creating two identical samples and showing one group (the treatment group) an ad that is not shown to another group (the control group), it is possible to attribute any observed opinion differences to the ad. Although the power of experiments is widely known within the advertising and market research testing community (Patzer 1996; Gardner and Belk 1980; Steckel et. al. 1991), because of the difficulty and cost associated with re-interviewing respondents, this community usually conducts "one shot" experiments (i.e., a single interview for each respondent). We add a new wrinkle to experimental design to make them more powerful. In short, we combine an experimental design with a panel framework.

A panel framework is a desirable improvement over a static design because respondents' opinions can be affected by both the fact that the respondent has seen an ad (i.e., the ad's effect), and the fact that the respondent saw the ad in the context of an interview session (i.e., the interview effect). Thus, comparing responses from a control group to a group exposed to an ad does not adequately isolate and measure the effect of the ad because two forces are simultaneously affecting a respondent's attitudes. The first of these two forces deals with the effect of viewing the ad (which is the effect that is desired). The second deals with the fact that a short-term "pushed" effect may exist because the respondent was *just* shown the ad – plausibly exaggerating the respondent's reaction to it.

An experimental panel design allows us to disentangle these two forces and determine whether any opinion change falsely attributed to the "treatment" actually results from the interview effect. This allows us to measure the rate of decay associated with an ad's effectiveness. In other words, by taking a baseline measurement prior to the ad test, and then following up the test with another measurement, we create another analytical dimension that can be used to assess an ad's effectiveness. By comparing the results to questions of interest both across time and across groups, it is possible to identify not only the salience of the ad (by comparing the responses from the ad test to those collected in the follow-up measurement), but also the extent that interview effects produce the observed results (by comparing the prior measurement to the responses from the ad test).

In this section we (briefly) articulated three aspects of the design we employ to test the effectiveness of ads. In addition, we outlined why each of the aspects represented an improvement over the existing methodology. We now turn to a

description of the respondents used in the study, as well how they were recruited by the research firm Knowledge Networks.[4)]

THE RESPONDENTS: KNOWLEDGE NETWORKS' METHODOLOGY

Knowledge Networks (KN) provided the Yale Advertising Study with the internet-enabled panel of respondents.[5)] To select its panelists, KN utilizes list-assisted RDD (Random Digit Dialing) sampling techniques on a quarterly updated sample frame consisting of the entire United States telephone population. KN excludes only those banks of telephone numbers that have zero directory-listed phone numbers (i.e., KN's telephone numbers are selected from "1+ banks"). All numbers have an equal probability of selection, and the sampling is done without replacement.

After generating the initial list of telephone numbers, the sample preparation system then excludes confirmed disconnected and non-residential telephone numbers. Next, the sample is screened to exclude numbers that are not in the Microsoft WebTV Network. This results in the exclusion of approximately 17% of the United States population.[6)] Telephone numbers for which KN is able to recover a valid postal address (about 50%) are sent an advance mailing informing them that they have been selected to participate in the KN Panel. In addition to containing some information about the KN panel, the mailing also contains a small monetary incentive to encourage cooperation.

Following the mailing, the telephone recruitment process begins. The numbers called by interviewers consist of all numbers sent an advance mailing, as well as 50% of the numbers not sent an advance mailing. During the recruitment interview, which typically requires about ten minutes, the interviewer informs the household member that they have been selected to join the KN panel, and that in return for completing a short survey weekly, the household will be given an interactive WebTV set-top box and free monthly internet access. All members in the household are then enumerated, and some initial demographic variables (i.e., age, gender, household mailing address) are collected. The household cooperation rate for the period examined averaged 56%.

The first survey that the panelist receives is a profile survey that collects standard demographic information. Panelists are not eligible for subsequent surveys until completing a profile survey. The first panelist joined the KN panel in June of 1999 and as of October 26, 2000, there were 43,448 respondents that had completed the initial profile and eligible to be randomly selected for our experiments.

The panel of Knowledge Networks (KN) is ideal for the Yale Advertising Study for four reasons. First, it is the only panel that harnesses the power and speed associated with internet surveys without abandoning a commitment to random sampling. Second, by providing all of its respondents with identical hardware (i.e., the WebTV) and free internet access, all respondents have a uniform survey-taking experience. Third, as described earlier, the WebTV's provided to the respondents by KN have several technological features that make the testing of multi-media content ideal. Fourth, since KN utilizes a panel design, re-interviews are unproblematic. As a result, Knowledge Networks' panel enabled us to easily implement the improvements noted in the section 'New Innovations'.

EXPERIMENTAL DESIGN: CONTROL AND TREATMENTS DEFINED

Our project draws upon the strength associated with a fully randomized, controlled experimental design. All of our experiments included randomly selected control and treatment groups from KN's panel of respondents. Treatment groups – those shown ads – were administered surveys addressing issues surrounding the 2000 U.S. presidential campaign with the ads embedded in the survey. Respondents were asked a host of questions about the ads themselves and were followed up with several attitudinal and behavioral questions dealing with the election. The interested reader can visit our website, www.yale.edu/newmedia, to see a screen captured example of an experiment and full details about the questions asked.

The ads used in our experimental design were for either Republican presidential candidate George W. Bush or Democratic presidential candidate Al Gore. The 30-second ads used were actual television commercials produced by either the candidates themselves or by the national party committees.

We devised two basic experimental designs. In the one political ad design, the experimental treatment consisted of showing the respondents a single political ad for a presidential candidate. While the one political ad design does not accurately represent the conditions of an election where candidates respond to each other's political ads, we included it because it allows for us to isolate the effects of a single campaign ad and it also allows us to avoid the confounded variables problem (Ansolabehere and Iyengar 1995: 23; see Clinton and Lapinski 2000 for a complete discussion). This design was used twice.

We also used a two ad design (i.e., respondents watched two political ads) in four experimental treatments. Within this paired arrangement, we varied the treatment groups to reflect several different scenarios. These scenarios allow us to investigate the advertising strategies of political candidates (which are

often conditional upon the strategies of opponents), by assessing the reaction of voters to information from competing sources (Ansolabehere and Iyengar 1995). We put a special emphasis on "negative" or "attack" ads because of their commonplace in American politics (Ansolabehere and Iyengar 1995; Iyengar and Prior 1999). Every effort was made to pick ads that were created and aired by the candidates themselves in approximately the same time period as the survey took place. Table 1 provides a breakdown of our experimental treatments along with their associated sample size.

Table 1
EXPERIMENTAL DESIGN

Respondent Group	*Type*	*Sample size*
Treatment A Experiment 1 (Bush Negative; Gore Negative)	Two Ad Design	516
Treatment B Experiment 1 (Gore Negative; Gore Positive)	Two Ad Design	603
Treatment C Experiment 1 (Gore Negative)	One Ad Design	671
Treatment D Experiment 1 (Gore Positive)	One Ad Design	688
Control Group Experiment 1		6,303
Treatment A Experiment 2 (Bush Positive; Gore Negative)	Two Ad Design	755
Treatment B Experiment 2 (Gore Negative; Bush Negative	Two Ad Design	719
Control Group Experiment 2		2,095
		Total N=12,350

ANALYTICAL RESULTS

Six different treatment groups over two experiments (two control groups, see Table 1 for details) comprising 12,350 respondents participated in the Yale Advertising Study. Our survey experiments were approximately 12-15 minutes in length, with the follow-up surveys after the experimental treatment lasting approximately five to seven minutes. The scope and size of the experimental design makes it impossible to provide anything but a few illustrative empirical findings from the project.[7] The findings reported have been selected as

illustrative of how our research design can be utilized to ascertain the effectiveness of television advertising. In so doing, we discuss the pitfalls encountered in trying to assess an ad's effectiveness, and how our design allows a researcher to accurately measure how well an ad achieves its intended effect.

The ultimate goal of advertising is to either directly or indirectly induce a certain behavior in its viewers. Consequently, the true measure of an ad's effectiveness is its ability to elicit or facilitate the desired behavior. In the context of study described in this paper, the behavior of interest is the respondent's vote in the 2000 U.S. presidential election.[8)] Thus, the assessment of a campaign ads' effectiveness must be based upon the ads ability to impact the viewers' voting decisions.

The experimental design that we employ permits two methods of measuring an ad's effectiveness: through direct self-assessment and indirect assessment. We first address direct sect-assessment.

Direct Self-Assessment

Direct self-assessment attempts to measure the effectiveness of an ad by directly asking the respondents to report the impact of an ad on the behavior of interest. Direct self-assessment usually takes two forms. The first form asks the respondent to evaluate an ad according to several criteria. The researcher then rates an ad's effectiveness by looking at the ad's "score" in those dimensions thought to be important in facilitating the desired behavioral response. The second, and more direct, form simply asks the respondent how effective the ad is in terms of inducing the desired behavior.

In order to illustrate these two forms of direct self-assessment, as well as related problems with this method, consider the problem of a candidate trying to decide whether to run a negative ad (i.e., an ad that mostly attacks the opponent) or a positive ad (i.e., an ad that mostly describes the sponsoring candidate's positions or desirable qualities). The first form of direct self-assessment would try to decide which type of ad is more effective based upon select criteria.

For example, one consideration might be the fear that a negative campaign ad may hurt the candidate running the ad due to the public's reported general dislike of negative advertising. We quantify this impact by examining the responses of 3,569 respondents who were asked, prior to either being shown any ads, or being informed that they were going to be shown any ads: "There is a lot of talk about negative campaigning these days. Would a negative campaign ad – one that pointed out negative things about an opponent – make you…" The results yield the response distribution reported in table 2.

Table 2
GENERAL ASSESSMENT OF THE EFFECT OF A NEGATIVE AD

Response Category	*Percent*
More likely to support the candidate running the negative ad	2
Not change your support	48
Less likely to support the candidate running the negative ad	31
Don't know	19

Note: Responses are based on the responses of 3,569 panelists collected during the second wave of video testing.

Table 2 provides evidence that candidates seeking the support of voters should *not* run negative ads, since negative advertising decreases their support among 31% of the population. The fact that this form of direct self-assessment relies on asking the respondents to rate the ads according to criteria of interest, it is critical to point out that this approach addresses the fundamental question of "what is the impact of either ad on the viewers' vote choice?" only tangentially. The reason for this is that it relies upon evaluative standards to proxy for the ad's actual effectiveness.

The second form of direct self-assessment judges an ad's effectiveness by using an experimental design and directly asking the respondents to assess the effectiveness of the ad. For example, this approach assesses the relative effectiveness of a negative Gore ad, and a positive Bush by showing the respective ad and then simply asking: "Still thinking about the ad you just saw, if the election were held today, would the *ad* make you…" The results are in table 3.

Examining the frequencies presented in table 3 reveals that no difference exists between the two ads – the percentage saying that they are more likely to vote for Bush after seeing a Bush positive ad (23 %) is statistically identical to the percentage saying they are more likely to vote for Gore after seeing a Gore negative ad (25%). Consequently, this experiment provides no evidence to support the conjecture that positive ads produce more support than negative ads.

Table 3
DIRECT SELF-ASSESSMENT OF AN AD'S EFFECTIVENESS (%)

	Gore Negative Ad				*Bush Positive Ad*			
	Dem.	***Ind.***	***Rep.***	***Tot.***	***Dem.***	***Ind.***	***Rep.***	***Tot.***
More likely to vote for GeorgeW. Bush	4	10	35	15	6	19	51	23
Have no effect on your vote	47	43	54	49	63	44	42	53
More likely to vote for Al Gore	41	14	8	25	24	4	2	13
Don't know	8	33	4	11	8	33	5	12

Note: *Results summarize the views of 729 respondents shown a Gore negative ad and 831 respondents shown a Bush positive ad.*

The two methods of direct self-assessment provide contradictory conclusions, as the results of table 3 directly contradict the results found in table 2. After being shown a negative ad by Al Gore attacking George W. Bush, only 15% said that they would be more likely to vote for George W. Bush (only 4% among Democrats). This is far from the 31% reported in table 2. In fact, this discrepancy can be further quantified by examining the behavior of the respondents who claimed that they would be "Less likely to support the candidate running the negative ad". Table 4 describes how the 195 respondents who answered such and saw the negative Gore ad compare to the 268 respondents who answered such and saw the Bush positive ad.

Table 4 confirms this contradiction. Despite previously reporting a negative electoral consequence for a candidate running a negative ad, no evidence exists to support such a claim. The support of the candidate not running the ad (i.e., Bush in the second column at 15%, and Gore in third column at 9%) is both: well below the 31% level reported in table 2, and statistically identical after accounting for sampling error.

The two methods of direct self-assessments examined in this section provide mixed results in terms of the ability to measure the effectiveness of ads. The first method of direct self-assessment – gauging an ad by its performance on criteria hypothesized to measure effectiveness – produces conclusions at odds with the second method of direct self-assessment – directly asking viewers how the ad affects the behavior of interest. This is a troubling contradiction because we would hope that alternative methods of measurement would provide us with the same findings to assure us of robust results. Given the

contradictory inferences revealed by the two methods of direct self-assessment, we turn to an alternative method of evaluating ads.

Table 4
EFFECT OF NEGATIVE AD EXPOSURE ON THOSE AVERSE TO NEGATIVE ADS

	"Less likely to support the candidate running the negative ad" and saw Gore negative ad	***"Less likely to support the candidate running the negative ad" and saw Bush positive ad***
More likely to vote for George W. Bush	15%	35%
Have no effect on your vote	48%	47%
More likely to vote for Al Gore	28%	9%
Don't know	8%	9%

Note: Table 4 summarizes the responses of the 195 respondents who saw the Gore negative ad on survey wave 2 and who answered "Less likely to support the candidate running the negative ad" in the question summarized in table 2 and the 268 respondents who saw the Bush positive ad on survey wave 2 and who answered "Less likely to support the candidate running the negative ad" in the question again summarized in table 2.

Indirect Self-Assessment

In the previous section, we measured an ad's effectiveness by directly asking respondents shown an ad to rate its effectiveness either in terms of a specific dimension of interest (e.g., believability) or in terms of a behavioral effect (e.g., presidential vote choice). However, there is reason to believe that *either* assessment is incorrect. The existence of "priming" effects, where a second product (or ad) that is evaluated is often evaluated differently than the first (i.e., systematically higher or lower scores) is well known. Priming occurs because the exposure (and cognition thereof) of the respondent to the evaluative environment affects the way that a respondent evaluates the project. For example, once the respondent is made aware of the fact that they are being asked to evaluate different products, the respondent may answer questions differently then they would normally (e.g., accentuating differences, responding in a socially desirable manner, or over-analyzing the question). A problem for any researcher is that although the existence of such effects is certainly plausible, it is impossible to understand the extent and magnitude of any such bias using a direct self-assessment, as any such assessment would

again be vulnerable to such a bias. Although our method is certainly an improvement over the standard procedures in that we do not need to rely on interviews conducted in a controlled environment (e.g., a interviewing facility), the possibility for the bias associated with any self-assessment still exists.

Our panel design allows us to account for this bias. Relying on responses given in different contexts on separate dates allows us to recover the extent to which the behavior of interest changes controlling for the different treatments (i.e., ads). In addition to asking a respondent to directly tell us about the ad's effectiveness, we also indirectly recover the effectiveness by comparing the response to a baseline response that does not contain any possible influence of the ad examined (shown in the previous experimental survey). In the panel experimental design employed in the Yale Advertising Study, two baselines were taken from which deviations resulting from ad exposure can be quantified. The first baseline comes from the presidential vote choice expressed in a survey prior to the ad experiment; and second, is taken from the presidential vote choice of the control group (i.e., shown no ads) in the ad experiment. The first baseline allows for us to control for respondent differences, but introduces possible error due to time dynamics, and the second baseline provides for the exact opposite. By using the two baselines simultaneously, the researcher can effectively control for both possibilities.

Table 5 contains the results of such an analysis. Respondents were asked who they would vote for three times during the presidential campaign: 1) prior to the ad test, 2) after being shown either a single ad (or both ads in the two ad design) or no ads (if they were in a control group), and 3) immediately following the November election.

Table 5
GORE SUPPORT OVER TIME BY POLITICAL PARTY AND AD EXPOSURE (%)

	No ad shown			***Gore positive ad only***			***Gore negative ad only***		
	Dem.	***Ind.***	***Rep.***	***Dem.***	***Ind.***	***Rep.***	***Dem.***	***Ind.***	***Rep.***
Gore prior	78	24	5	77	23	4	72	17	5
Gore after	83	43	10	83	51	10	82	36	11
Gore vote	84	44	10	87	51	10	82	47	11

Note: The "prior", "after", and "vote" reflect the responses to three different surveys. The table summarizes the responses of 328 respondents who saw only a Gore negative ad, 331 respondents who saw only a Gore positive ad, and 3124 respondents who never saw an ad.

Let us first consider the responses of a Democratic respondent not shown an ad (i.e., the first column of data). Prior to the ad test, 78% of the Democrats not shown an ad indicate an intention to vote for Al Gore. During the ad test (on which they did not see an ad), this percentage rises to 83%. On Election Day, this rises to 84%.

First, let us think about an analysis using the initial baseline (i.e., the prior responses of those shown an ad). In order to determine an ad's effect, we compare support levels prior to seeing the ad with support levels following the ad. For example, among independents, Al Gore's support rises from 23% to 51% among those who saw only the positive ad (28% gain). Among those independents who saw only the negative ad, the gain was only 19% (17% to 36%). This analysis – comparing the behavior before and after being shown the ad – is, of course, incomplete in that other factors could explain the observed shift (e.g., the political campaign itself).

We can account for this by comparing the results to a second baseline. The second baseline results from a control group interviewed at the same time, but shown no ads. For example, Gore's support among independents not shown an ad increased from 24% to 43% (19% gain). The 19% gain among independents never exposed to the negative ad is identical to the Gore's 19% gain among independents exposed to a negative ad. As a result, the correct inference is not that the negative ad produces a 19% gain in support, but rather, after controlling for external factors, the negative ad has absolutely *no effect* among independents. The effect of a positive ad is similarly tempered, as 19% of the 28% observed gain can be accounted for by forces other than the ad. Thus, the effectiveness of this positive Gore ad is not + 28 %, but instead closer to + 9%.

To illustrate the importance of using indirect assessments, consider what a direct self-assessment would reveal about the comparative effectiveness of the Gore negative and positive ad. Reproducing the analysis presented in table 3 for the respondents summarized in table 5, we arrive at the following (see table 6.).

Examining the relative responses of independents, we find that 27% of the independents exposed to Gore's positive ad and 19% exposed to Gore's negative ad reported that they would be more likely to vote for Gore. As previously outlined, both of these magnitudes are much larger than the effect revealed by the indirect assessments. In fact, we know that since Gore gained the support of 19% of independents not exposed to any ads, the 19% gain reported in table 6, is actually completely illusory once we control for external forces.

Table 6
DIRECT ASSESSMENT OF AD EFFECTIVENESS (%)

	Gore Negative Ad				*Gore Positive Ad*			
	Dem.	***Ind.***	***Rep.***	***Total***	***Dem.***	***Ind.***	***Rep.***	***Total***
More likely to vote for George W. Bush	3	10	37	18	3	11	25	13
Have no effect on your vote	60	70	54	60	43	62	68	57
More likely to vote for Al Gore	37	19	8	23	54	27	7	30

Note: Results summarize the views of 326 respondents who saw only the Gore negative ad, and 329 respondents who saw only the Gore positive ad. These respondents are the same respondents summarized in table 5.

The indirect analysis (and the experimental design that we employed that enabled the indirect analysis) allowed us to quantify the bias associated with direct self-assessment. Absent the indirect assessment, the extent and magnitude of bias associated with direct self-assessment would remain hidden. This would greatly hinder our ability to determine the true effectiveness (or lack of effectiveness) of an ad.

The final piece of analysis we perform focuses on the salience of the ads. In other words, how lasting of an impact does the ad have on its viewers? Although the context that we examine is less that ideal for this analysis (given the presence of an ongoing, continuous campaign), it is still possible to perform an analysis illustrative of the kind that is required. To examine the extent to which the support manifest in the ad experiment is lasting, we utilize the panel framework and simply compare the responses to the vote question on the ad test to the vote question on the post-election survey. Recalling the results in table 5, this exercise consists of comparing the bottom row to preceding row. As is immediately evident, the percentages are identical across the two rows – with the exception of the independents shown the negative Gore ad. This implies that whatever impact the ads produced on the voting decision of the respondents was sustained through Election Day.

The results in this section have shown how an experimental panel design, and specifically indirect assessment, can be used to accurately assess the effectiveness of an ad. By showing the different conclusions that direct self-assessment and indirect assessment yield, it is clear that the determination of an ad's effectiveness is far from straightforward. As a result, the researcher should utilize as many analytical tools as possible, as well as operate in the

most favorable (i.e., unbiased) conditions possible to maximize the probability that the assessments offered will be correct. The illustrative analysis presented in this section, although clearly intended as an overview only, are suggestive of the ways in which the experimental panel design that we utilize can be incredibly useful.

CONCLUSION

Our paper provides a novel approach to testing the effectiveness of ads not previously possible because of technological limitations. The Yale Advertising Study has provided empirical evidence dealing with the complexities faced by researchers in trying to assess an ad's effectiveness. Three conclusions are immediate from our analysis.

- Analysis relying on constructed criteria can be no better than the criteria themselves. Furthermore, nothing constrains the criteria to be an accurate measurement of an ad's effectiveness.
- Direct self-assessment, although arguably better than criteria-based analysis, is subject to unknown response bias. Given this bias, it is impossible to know how meaningful the assessments are.
- Indirect self-assessment through a panel design and a control group enables the researcher to directly quantify the effect of the ad in terms of the desired behavior. This makes it possible to control for (and quantifies) the bias resulting from direct self-assessment.

We hope that researchers benefit from our research, and extend it. Further analysis on this topic is currently underway by the authors.

FOOTNOTES

1. The exact dollar amounts for 1999 are as follows: broadcast television, $40,011 (millions); cable television, $10,429 (millions); radio, $17,215 (millions); internet, $1,940 (millions); newspapers, $46,648 (millions); magazines, $11,433 (millions). Source: Advertising Age, prepared by Robert J. Coen and McCann-Erickson Worldwide.
2. Internal validity focuses on mistakes that occur in measures (i.e., the extent to which the measurement itself is free of error). External validity focuses on what the measures actually measure (i.e., the extent to which that what is intended to be measured is actually measured). In other words, internal validity is the extent to which changes in a dependent variable are actually caused by changes in an independent variable, and not an extraneous variable. External validity is the extent to which the measures (i.e., collected data) in a particular research project can be generalized. (Patzer 1996: 43)

3. Respondents must meet four other conditions to be included in the sample. These four conditions include 1) ability to speak English, although it need not be there first language; 2) own at least one television set; 3) live within an area that received WebTV coverage; 4) over the age of eighteen.
4. Knowledge Networks (http://www.knowledgenetworks.com) is headquartered in Menlo Park, CA.
5. Professor John Lapinski is grateful for the financial support of the Institution for Social and Policy Studies (ISPS) at Yale University. ISPS provided the primary financial support for the Yale Advertising Study.
6. This number reflects KN's panel at the time our experiment were conducted. We were in the field from September 2000 to November 2000.
7. Those interested in the effectiveness of political advertising should consult the authors' substantive paper (Clinton and Lapinski 2000).
8. A host of questions associated with voting behavior were assessed in the Yale Advertising Study. See our website www.yale.edu/newmedia for a detailed set of questions.

REFERENCES

Ansolabehere, Steven and Iyengar, Shanto 1995. *Going Negative: How Political Advertisements Shrink & Polarize the Electorate*. New York, NY: The Free Press.

Batais, Laurent and Spitzer, Laurent. 2000. The Effectiveness of Television Advertising in France. *International Advertising: Realities and Myths*: Thousand Oaks, CA: Sage.

Gardner, David M. and Belk, Russell W. *A Basic Bibliography on Experimental Design in Marketing*. Chicago, IL: American Marketing Association.

Helgesen, Thorolf and Micalsen, Morten. 2000. Short-Term Advertising Strength: New Empirical Evidence from Norway. *International Advertising: Realities and Myths.* Thousand Oaks, CA: Sage.

Iyengar, Shanto and Prior, Marcus. 1999. Political Advertising: What Effect on Commercial Advertisers. Working Paper.

Jones, John Philip. 1995. *When Ads Work: New Proof that Advertising Triggers Sales.* New York, NY: The Free Press.

Kish, Leslie. 1995. *Survey Sampling*. New York, NY: John Wiley & Sons.

Lohr, Sharon. 1998. *Sampling: Design and Analysis*. New York, NY: Brooks & Cole Publishing.

McDonald, Colin. 2000. How Single-Source Research First Developed. *International Advertising: Realities and Myths*. Thousand Oaks, CA: Sage.

Ogilvy, David. 1985. *Ogilvy on Advertising*. New York, NY: Random House.

Patzer, Gordon. 1996. *Experiment-Research Methodology in Marketing*. Westport, CT: Greenwood Publishing Group

Steckel, Joel H., DeSarbo, Wayne S. and Mahajan, Vijay. 1991. On the Creation of Acceptable Conjoint Analysis Experiment Designs. *Decision Sciences,* 22: 435-442.

THE AUTHORS

Joshua Clinton is Statistical Research Scientist, Knowledge Networks, United States.

John S. Lapinski is Assistant Professor, Department of Political Science; Director, New Media Workshop; and Research Fellow, Institution for Social and Policy Studies and Center for Internet Studies; Yale University, United States.