Methodological Issues in Advertising Laboratory Experiments

Lars Bergkvist, Marketing Research Innovation Centre, University of Wollongong, Australia
John R. Rossiter, Marketing Research Innovation Centre, University of Wollongong, Australia

This paper discusses seven factors that limit the validity and generalizability of advertising laboratory experiments. These factors include exposing ads only once to participants, which limits the validity of the results, and the use of a single ad or brand in the study, which limits generalizability. Examples are offered of studies limited by the factors discussed in the paper. The paper concludes with recommendations of how future advertising laboratory experiments can be improved.

[to cite]:

[url]:
http://www.acrwebsite.org/volumes/12981/volumes/ap07/AP-07

[copyright notice]:
This work is copyrighted by The Association for Consumer Research. For permission to copy or use this work in whole or in part, please contact the Copyright Clearance Center at http://www.copyright.com/.
METHODOLOGICAL ISSUES IN ADVERTISING LABORATORY EXPERIMENTS
Lars Bergkvist, Marketing Research Innovation Centre, University of Wollongong
John R. Rossiter, Marketing Research Innovation Centre, University of Wollongong

ABSTRACT
This paper discusses seven factors that limit the validity and generalizability of advertising laboratory experiments. These factors include exposing ads only once to participants, which limits the validity of the results, and the use of a single ad or brand in the study, which limits generalizability. Examples are offered of studies limited by the factors discussed in the paper. The paper concludes with recommendations of how future advertising laboratory experiments can be improved.

INTRODUCTION
Laboratory experiments are the most common method in academic advertising research and the method is enjoying increasing popularity. In 1984, eight out of 30 articles (27%) in the Journal of Advertising were based on advertising experiments but, by 2004, the share had more than doubled: 15 out of 26 articles (58%) in JA were based on experiments. The main advantage of experiments is high internal validity: they allow for control of important factors influencing the effects of advertising, for example competitive interference or exposure frequency, while other factors, for example, benefit claims or spokespeople, can be varied between experimental groups and causal inferences made. A disadvantage is that controlling for factors extraneous to the problem under study may limit external validity to the extent that the results have no value for understanding and explaining real-world advertising phenomena. Lack of realism has received attention from researchers and several articles discussing limitations with advertising experiments and consumer behavior experiments have been published (e.g., Bogart and Lehman 1983; McQuarrie 1998; Wells 1993). However, lack of realism is not the only potential problem with advertising experiments. Many advertising experiments have other limitations in their design that limit the value of the studies. These limitations can, for example, relate to the generalizability of the results. Generalizability limitations have received little attention among scholars, even if they are at least as important as limitations due to lack of realism.

This paper reviews and discusses limitations in academic advertising experiments and recommends how they can be remedied in future studies.

FACTORS LIMITING THE EXTERNAL VALIDITY OF ADVERTISING EXPERIMENTS
In the following sections five factors that limit the external validity, or realism, of advertising experiments are discussed. In addition one factor that limit the generalizability and one factor that limits the value of experimental results are discussed. These factors have been selected as they are believed to unnecessarily reduce the value of advertising experiments. All of these factors can and should be addressed when advertising experiments are designed. Although there are other factors that limit the realism of advertising experiments (see, e.g., McQuarrie 1998), it is not always possible to address these without harming the internal validity of the experiment. These other factors have not been addressed in this paper.

Single exposure
In most advertising experiments participants are exposed once to the ad or ads in the study. McQuarrie (1998) reports that 84% of advertising experiments published in leading academic journals (JAR, JA, JCR, JM, and JMR) between 1990 and 1997 used a single exposure.

The minimum effective frequency, the number of exposures at which an ad achieves its communication objectives, varies between product categories, situations, and target groups, but it is rarely a frequency of one, even for a direct-response ad (Rossiter and Bellman 2005). This means that advertising experiments with a single exposure may underestimate the effects of advertising. Even if processing of ads in advertising experiments is more focused and deeper than in real life, it is by no means certain that a single exposure will be sufficient for the ad to achieve its communication objectives. This applies in particular to TV ads, which participants cannot process at their own pace. Thus, small or non-existent effects of advertising may be the result of insufficient processing and not of ineffective ads in the experiment. For the same reason, it may well happen that an experiment fails to produce differences between experimental groups that would have occurred with multiple exposures. It may also happen that use of a single exposure leads to differences between experimental groups that would not occur with multiple exposures, since insufficient processing may mean that participants do not have sufficient grounds to form an overall impression of a brand but focus on certain details. An illustration of the importance of frequency can be found in a study by Berger and Mitchell (1989) who found that attitude accessibility and confidence was significantly higher, and the attitude-behavior relationship stronger, for participants exposed to the same ad three or four times compared to participants only exposed once.

Immediate measurement
Except for direct-response ads, there is generally a delay between exposure to advertising and brand choice. However, most academic advertising experiments are carried out with measurement of dependent variables immediately following exposure. In fact, 95% of experiments published in leading journals between 1990 and 1997 measured the outcome variables immediately following exposure (McQuarrie 1998). Immediate measurement means that the information in the ad is easy accessible in memory, while delayed measurement increases the likelihood that information decays or is naturally interfered with by other ads (Keller 1987). It has, for example, been shown that brand attribute judgments become less positive over time and that consumers rely more on specific product attributes when forming purchase intentions in immediate measurement and more on general product category attributes when forming intentions in delayed measurement (Mazursky 1990). Another illustration of the difference between immediate and delayed measurement can be found in Bergkvist and Rossiter (2005). Their study showed that Attitude toward the Ad (A_ads) influenced Brand Attitude (A_brands) when both were measured immediately following exposure to the ad. However, there was no relationship between the A_ads measure taken in immediate measurement and the A_brands measure taken in delayed measurement, which shows that the relationship between the variables is different depending on when measurement is made.

Measurement of brand awareness not valid
Almost all academic advertising experiments measure brand awareness communication objectives wrongly (Rossiter and Bellman 2005; Rossiter and Percy 1997). Most experiments measure recall or recognition of the brand from the ad, rather than to measure it independently of the ad, which is how brand awareness occurs in the real world. If brands in the product category are typically chosen by brand recall, then a delayed, category-cued measure of brand name recall should be used; if brands are typically chosen by recognition of the pack or logo, as for most fmcg brands at the point-of-purchase, then a competitive, speeded recognition test should be used. Brand “awareness” results based on brand name recall (“unaided”) or brand name recognition (“aided”) from the ad-testing context itself bear no systematic or valid relationship to independent-of-context measures. Achievement of valid brand name recall or valid brand recognition, as appropriate, is essential for subsequent brand evaluation measures, such as brand attitude or brand purchase intention, to be meaningfully interpreted.

Unrealistic ads

Some advertising experiments use real-world ads, for example, real ads for real brands taken from another market that are unknown to the participants (e.g., Baker, Honea, and Russell 2004; Bergkvist and Rossiter 2005). However, it is also common to use ads that have been designed by the academics doing the research. These latter ads often are amateurish and disturbingly lacking in realism. For example, Craig and Shimp (1990) ran an experiment in which they used two home-made print ads for the same brand of low-alcohol beer. The ads were designed to be identical except for the experimental variation in the text and the picture. The headline in both ads read “Introducing the new breakthrough in beer” and the layout of the ads was dominated by text (in fact, there were between 130 and 150 words in each ad, excluding the headline). The text in the ads was written in an academic style, for example, one of the ads contained the lines “Unlike regular beer, Break’s lower alcohol and lower calorie content allow you to have a great tasting beer while keeping physically and mentally fit.” Moreover, part of the experiment was to vary the attractiveness of the couple used as spokespeople in the ad. This meant that the picture in one of the ads was chosen because the couple in it looked unattractive according to pre-tests! In short, the ads were nothing like real-world beer ads, which tend to be dominated by attractive pictures, have short headlines, little or no text, and often are witty or humorous. A second example of unrealistic print ads can be found in the study by Schumann, Petty, and Clemons (1990) which was an experiment with eight different ads for a pen. The two sample ads shown in the article are glaringly amateurish, with static layout, typewriter typeface, academic style of writing, a long key benefit claim placed at the top of the ad immediately below the headline, no brand logo, and so forth. Their ads have virtually no similarity to real-world ads and any findings based on them lack external validity.

The problem with using home-made ads is that mental processing of these ads is likely to differ from the processing of real-world ads. If ads are badly laid out, use home-made pictures, or the text is dry and boring, participants are not likely to process the ads in the same way as they would with real-world ads. In many product categories, for example, soft drinks, beer, or spirits, most advertising comes in the form of image advertising which may do nothing more than show typical or aspirational users of the brand or convey a single abstract benefit such as “cool” (Rossiter and Bellman 2005; Sutherland and Sylvester 2000). Now, if academic advertising experiments employ dull home-made ads with large blocks of text, large number of benefits, and small pictures in these product categories, it is likely that the participants will process them rationally, based on the verbal arguments, instead of emotionally as they would normally do with ads in such product categories. If the focus of the experiment is on testing details of the ads, as is the case with the research within the Elaboration Likelihood Model paradigm (Petty, Cacioppo, and Schumann 1983), for example, testing “central” versus “peripheral” verbal arguments then effects are likely to be increased if the ad is dominated by text instead of pictures, especially if the ad is for a brand in a product category where image advertising dominates. It is also likely that differences between ads will occur on the “negative” side of the dependent variable, particularly brand attitude, with poor ads – another limitation of external validity because such weak ads would never be approved for real campaigns. Another problem with amateurish ads is very low purchase intention scores.

Artificially induced response sets

It is common in advertising experiments to set up experimental conditions that influence how participants process the information in the ads. For example, in a typical experiment Goodstein (1993) informed participants before they were exposed to the ads either that they would be asked about their impressions of the ad or their impressions of the brand, thus focusing their processing on either of the two. Inevitably, instructions of this type influence how participants process the ads and, as a consequence, the experimental results. Indeed, in most cases it is part of the experiment to test the effect on outcome variables of different instructions. The question is what the value of these results is. It is not likely that advertisers would be able to instruct their target audience how they should process the advertising they are exposed to. (Try to imagine a TV ad starting with an instruction to focus on the brand, not the ad, during the following 30 seconds.)

Just one or two ads

It is common practice to include only one or two ads in advertising experiments. This raises the issue of whether you can generalise to all advertising or to a certain category of advertising on the basis of one or two ads (Wells 2001). There are numerous differences between ads, even within the same category, as illustrated by the following quote:

“Humorous advertisements differ from each other in many ways. Some are attractive and some are unattractive. Some are gentle. Some are clever. Some are funny and some are not so funny. Some poke fun at men, or women, or animals, and some do not. Some are for new brands and some are for established brands. Some are for funny products and some are for serious products. Some are for services. Some are in newspapers. Some are in magazines. Some are on the radio. Some are on television. In short, the category is heterogeneous.” (Wells 2001, 494)

Furthermore, there are also differences in the mental processing of, and the responses to, advertising between product categories and between brands within categories. Product categories and brands are expected to vary with respect to the mental processing of advertising depending on the level of involvement (low or high) and type of purchase motive (informational or transformational), that is, the two dimensions in the Rossiter-Percy-Bellman Grid (Rossiter and Bellman 2005; Rossiter and Percy 1997).

The differences between ads and between product categories and brands means that, in effect, an advertising experiment with only one ad or brand tries to generalise to a heterogeneous population on the basis of a sample of one (Wells 2001). This means that results from an experiment with, for example, a painkiller ad (a typical low-involvement informational product) may very well not hold
for an ad for a vacation trip (a typical high-involvement transformational product). In particular, highly likable ads are a requirement in the low-involvement-transformational cell but not in the other three cells (Rossiter and Bellman 2005; Rossiter and Percy 1997).

**Small differences between experimental groups**

Advertising experiments, like other types of consumer behavior experiments, often result in small differences between experimental groups. One reason for this is that researchers tend to focus on statistical significance rather than effect size (Peterson, Albaum, and Beltramini 1985). An example of this can be found in Muehling and Sprott (2004) who reported statistically significant ($p < 0.05$) differences between experimental groups for Attitude toward the Ad ($A_{Ad}$) and Brand Attitude ($A_{Brand}$). However, the size of the differences was less impressive: For $A_{Ad}$ the group mean scores were 5.68 and 5.29, respectively, and for $A_{Brand}$ 5.91 and 5.58. (The authors do not report the length of their answer scales, but presumably they used 1 to 7, 7-point scales.) This means that both experimental groups had clearly favorable ad and brand attitudes, even if there were minor differences between groups.

This tendency to focus on small but statistically significant differences should be contrasted with how advertising practitioners use advertising copy tests. It is not likely that an ad would be rejected on the grounds that the mean brand attitude score was .33 lower, on a 7-point scale, for one ad than for another. Instead, practitioners tend to look at the share of respondents for the different response categories and reject ads if the percentage that dislikes an ad is too high or if the percentage that likes an ad is too low.

**SUMMARY AND RECOMMENDATIONS**

This paper addressed seven factors that limit the value of advertising experiments. Most of these factors unnecessarily limit the external validity and thereby the applicability to the real world of the experiment, but two of the factors limit the value of the experiment in other ways. The first of these two, the use of just one ad, or perhaps two ads, in the experiment, limits the generalizability of the experiment. A single ad or even two ads does not offer grounds to generalise findings across different types of ads or different types of product categories. The second factor, small differences between experimental groups, does not limit the realism or the generalizability of the experiment, it simply makes the experiment uninteresting from an academic as well as a practitioner point-of-view.

It can be argued that the five factors that limit the realism of advertising experiments (single exposure, immediate measurement, unrealistic ads, irrelevant product categories, and artificially induced response sets) are necessary in order to control for extraneous factors. After all, the whole point of an experiment is to control for as many factors as possible at the same time as the factors under study are varied between the experimental groups. However, to improve the realism of the five factors discussed in this paper would not have a negative effect on the experimental control, or internal validity, provided by advertising experiments. Using multiple exposures or several ads would not lessen experimental control, even if it does mean somewhat increased effort for the researcher. Using real ads for real products means that an extra effort is required to find suitable ads, but it does still allow for experimental control.

Other authors have discussed additional realism factors that have not been addressed in this paper. For example, McQuarrie (1998) mentions that ads should be embedded and not shown separately, that the outcome variables should be brand choice, that experiments should include competitive interference, and that experiments should include ads for familiar brands and not only new brands. These factors have not been addressed in this paper because it is believed that they cannot be addressed without jeopardizing experimental control or because they may entail practical difficulties of a large number of conditions, as well as sampling problems of their own. Also, if experiments are run with ads for familiar brands the effects of previous knowledge, brand attitude, and so forth, has to be accounted for in the experiment, which means that a larger sample of participants will be needed.

Based on the discussion in this paper, the following recommendations are offered for future advertising experiments.

*Advertising experiments should use multiple exposures to the ads in the study.* The effective frequency varies between situations but is generally greater than one. The number of exposures should be set depending on whether it is print, radio or TV advertising, whether the target audience is new category users, and so forth (see Rossiter and Bellman, 2005, for an overview of factors influencing effective frequency). It is also recommended that exposures be spread in time in order to simulate a real-world campaign. Most advertising campaigns run for at least a week and sometimes up to three or four weeks and ideally repeated exposures should be spread in time accordingly.

*Measures of brand awareness should be brand based.* Recognition or recall of the brand, as appropriate, should be measured independent of the advertising experiment context.

*Measurement of outcome variables should be made after a delay.* The effects of advertising should not be measured immediately following exposure, but after a realistic delay. The length of the delay should vary according to the average interval in the product category between advertising exposure and purchase opportunity. An absolute minimum would be a few hours with some interfering tasks between exposure and measurement.

*A minimum of four different types of ads should be included in the study.* It may not always be possible to include several different types of ads in the same experiment, but it is always possible to repeat an experiment using different ads. It is recommended that ads be chosen to represent the four quadrants in the Rossiter-Percy-Bellman Grid (Rossiter and Bellman 2005; Rossiter and Percy 1997).

*The ads in the experiment should be realistic.* A large majority of academic marketing researchers do not have the same talents that professional copy writers and art directors have. Therefore, it is better to ask advertising professionals to design the ads that are used in advertising experiments or to use ads taken from other markets that are unknown to the participants in order to get ads that are realistic and of a persuasive quality likely to be effective in a real campaign.

*The response of participants should not be induced in any way.* Participants in experiments should not be told how to process the information in the ads, only instructed to look at the ads as they normally do.

*Practical significance of results should be considered in addition to statistical significance.* Not all statistically significant results are practically significant or interesting and statistical significance should never be the single criterion for presenting a result.

**REFERENCES**


Confidence, and the Attitude-Behavior Relationship,” *Journal of Consumer Research*, 16 (December), 269-79.


