Principles of Measuring Advertising Effectiveness

By David Olson

Testing the effectiveness of advertising is a big business.

With the cost of producing a commercial at $500,000 or more -- and media budgets of $50 million -- there is a natural desire, on the part of advertisers, to make sure that the advertising they are running is effective.

John Wanamaker's famous quote, "I know that half of my advertising money is wasted... I just don't know which half," practically begs for a method to identify effective ad campaigns, commercials, or print ads.

Research companies, recognizing this need, have developed a myriad of testing approaches. They use conceptual models of how advertising works, various methods of gathering the data, different response measures, and different analytic approaches.

It's safe to say that there is no general consensus, on the part of advertisers or advertising researchers, as to the best way to test advertising.

This tutorial will lay out a taxonomy of the different approaches used, and some of the strengths and weaknesses of each approach. It will describe the decisions that an advertiser must make in choosing the most appropriate method for their particular type of product or service, their customers or users, and their type of advertising.

A Taxonomy of Advertising Testing Approaches

The advertising testing approach you should use depends on:

- The type of advertising (TV spot, print ad, etc.)
- The type of stimulus you want to test (storyboard, finished commercial, etc.)
- The measures of effectiveness you want to obtain.

Type of Advertising

The more expensive the advertising is to produce, the more compelling the economic argument is to test that advertising before running it. As a result, most advertising testing is done in regards to television commercials.

Print ads are also frequently tested, though not as often. More rarely, radio ads, outdoor (billboards), and even Website banner ads will be tested too. Direct mail is tested, but by small-batch mailings where the evaluation is based on direct response measures. The same can be done with web site banner ads that have a click-through response feature.

Most of the issues discussed below relate to testing TV commercials. It is the predominate type of ad testing done. It also tends to be a more difficult task for the researcher.

Type of Stimulus

In order to test an ad, you obviously have to create some sort of stimulus to expose to respondents. The validity and accuracy of their response is only as good as the stimulus.

Clearly the most valid stimulus is the advertising in its final, finished form. For simple print ads, this is
not much of a problem. For expensive TV commercials, however, only testing the ad after it has been produced largely defeats the purpose of doing ad testing in the first place.

In order, from roughest to most finished form, these are the types of stimuli you might want to test:

1. **Storyboards.** These are the actual hand-drawn storyboards the agency uses internally to evaluate ad ideas, and which the agency uses to expose the proposed ideas to the client. They give only a very rough idea of what the final ad might look like. Sometimes a recorded track of the spot's music and words is played, as the agency creative points to each frame to communicate the action and flow of events.

   Testing with storyboards is typically done for very early qualitative testing using focus groups. The agency will present the idea just as it did for the client.

   This method of exposure is completely artificial. This is, in fact, one of its main virtues. As the ad stimulus gets more refined, respondents start to believe that the final ad itself is being evaluated.

2. **Roughs.** This is the form of stimulus used most often in advertising testing. The agency will produce a prototype of what the final commercial or ad might look like, but with primitive production values.

   It is exposed to consumers just as a finished ad might be. If the final ad is a 30-second spot, the rough video will be about 30 seconds long as well.

   There are several variants of roughs that are used:

   a) Animatics are cartoon drawings - essentially just the storyboard scenes or frames put on video with a narrative track and music bed, and with only very primitive animation effects.

   b) Photomatics are like animatics but will use some photographic scenes taken from stock footage, to enhance this stimulus' production values.

   c) Steal-o-matics will use actual video footage taken from other sources such as prior finished commercials. This makes the prototypes more realistic.

   For testing rough print ads, prototypes often are created with stock photographs and in-house graphics.

   Producing these rough prototypes is not cheap -- sometimes they can cost $25,000 and more – but they can be critical to getting a decent read from consumers.

3. **Finished Ads.** Showing consumers the actual finished ad will obviously yield the most valid results. But as mentioned above, it's rare that ad testing is done with finished ads, due to the economics involved.

**Interviewing Methodologies**

Advertising testing employs a variety of research or interviewing methods.

**Qualitative tests**

In some cases, focus groups, or very small samples of individual personal interviews, are used to get some reaction to advertising prototypes.

The advantage of these tests is they are inexpensive, and they yield rich diagnostic information. However, because they are such small samples, and in the case of focus groups are subject to bias due to group effects, they are generally of limited value in testing advertising.

**Quantitative tests**

- **One-on-Ones/Personal Interviews.** In these tests, consumers are shown the advertising stimulus in a personal interview. Respondents may be recruited at a shopping mall, or interviewed in their homes, or in other ways. Sometimes they are shown the ad in a "clutter reel" among
other ads, or they see the test ad alone.

Procter & Gamble refers to such tests as "C&R tests," (comprehension and reaction). They measure:

- Does the viewer understand what is going on in the commercial?
- What sort of feelings or attitudes does the ad generate, in this "forced-viewing" situation?
- **Theater Testing.** In these tests, a large number of people (50 to 250 or more depending on the test) are recruited to attend a two- or three-hour research session.

During the course of the session, they are exposed to a number of test commercials. In many cases, the commercials are shown in the context of a complete TV show (usually an old rejected pilot). They then fill out a questionnaire measuring their recall of the ads they were shown, and other questions.

This kind of testing is less artificial than one-on-one interviews since the ads are exposed within a TV-program environment. However, people know they are taking part in a research study and thus are responding in the role of "professional respondents."

- **In-Home Testing.** In these tests, advertising prototypes are shown to people in their homes, in the context of their regular TV viewing. This is obviously the most realistic way to expose an ad, but it is fraught with other problems.

During the 1960s and 70s, this was the way most advertising was tested. However, analyses showed that these tests failed to control for crucial variables that influence results. The TV program the ad appeared with had a huge effect on the commercial's recall scores.

These tests were, in essence, "too realistic." A more tightly controlled, though unfortunately more artificial, testing environment was needed to give reliable results.

The growth of cable TV in the 70s and 80s permitted testers to cut different ads to different households and read the sales effects. IRI's BehaviorScan method was the first to set up these "real-world" tests.

**Measures of Effectiveness**

Following are the primary types of effectiveness measures:

- **Recall.** Some tests will provide a "recall score" for the commercial. This measures how many viewers were able to spontaneously recall the ad when questioned later.

  The recall score has many detractors. Agencies often argue that an ad can be effective even if people cannot easily recall it. And recall is highly influenced by such things as the brand name, so a comparison of scores across brands is problematic.

- **Recognition.** These measures are similar to recall, except they show people the commercial later and ask if they remember having seen it earlier.

  This technique still assumes that memory of an ad is an important predictor of effectiveness.

- **Narrative Playback.** Some tests ask respondents to recall the narrative or storyline.

- **Attitudes Toward the Ad.** These questions measure how people felt about the ad they saw. Some of these measures include: likeability, stimulation or entertainment, relevance, newsworthiness, uniqueness, believability.

- **Persuasion.** Advertising agencies argue that the real purpose of an ad is not to be remembered, but to motivate purchase. Market research companies developed a variety of proprietary methods to gauge the actual persuasiveness of the ad.

One limitation of these measures is the question of whether an ad, with a single exposure, can really motivate someone to buy a brand -- especially a well-known mass-market brand.
Exotic Measures. Beyond the measures above, ad testing methods sometimes have included:

- **Moment-by-Moment** assessment of the commercials, where respondents continuously turn dials to show how much they like the ad at each point during its broadcast.
- **Eye-Tracking** which measures what parts of the print ad the respondent is looking at, using special cameras.
- **Autonomic Measures** such as brain wave patterns or galvanic skin response to measure arousal and involvement.

Which Measures Are Best?

Given the wide range of measures available, which should an advertiser use for advertising testing?

Choose a method that is most appropriate for the product category and for the type of advertising being considered. For a new packaged good product, generating awareness and trial of that product is the primary goal of the advertising. In this case, recognition and persuasion may be the most relevant metrics.

For a well-known icon brand, measuring the likeability and relevance of the ad may be most appropriate. For a knowledge-rich brand like a new electronics product, understanding and playback of key sales points may be most critical, following multiple ad exposures.

The Advertising Research Foundation (ARF) conducted a comprehensive evaluation study to assess measures of ad effectiveness by validating them against BehaviorScan split-cable test market results.

Their surprising conclusion was that the single best predictor of post-advertising purchase of a brand among the measures listed above was "Likeability". If many respondents said they "liked" the ad, it had a higher probability of generating purchases.

Of course, this raises the question of what it means when someone "likes" an ad. Does it mean it was funny and entertaining? Did it tell them something important?

Measurement Objections

Whatever the technique, interpreting results brings up several recurring issues from advertising agencies.

- **Norms.** Every test compares the results to some sort of standard or "norm." The question will be whether that norm is fair, or if it is too high.

  If the test brand is small and not well liked, its ads may consistently fall below overall category norms even if the advertising is excellent. If it is a new product, it will have a difficult time matching norms based on strong, well-known brands.

- **Single vs. Multiple Exposures.** Most ad tests involve a single exposure (or at most, two exposures) to the ad. In the real world, ads are seen multiple times. The agency will question whether a single exposure is sufficient to gauge people's reaction. They are especially concerned about subtle, or complex types of ads that may flunk a single-exposure test, yet do very well when they run on the air.

- **Sample Composition.** Most market research companies want to define their sample broadly, since it is much cheaper to recruit a general sample than a specifically defined one. But advertising -- even mass media TV advertising -- is targeted to segments.

- **Decision Criteria.** Some clients will want to establish go/no-go decision rules regarding the ad results based on the key measures. They may want the ad to be above average for all ads tested by that client before production. The agency will argue about the fairness of such yardsticks given the uncertainties of the measurement techniques.

At the end of the day, the question remains: Which technique should one use to evaluate advertising?
The key considerations are:

1. **Your Conceptual Model of How Advertising Works.** Pick a technique that is consistent with what you think is important. If communicating solid facts about products is key to your advertising, then measures of comprehension will be important. If your brand requires emotion and imagery, then more qualitative techniques may be most appropriate.

2. **The Reliability and Validity of the Technique.** If the technique’s results bounce around due to random factors -- if its test-retest reliability is poor -- then you are reading tea leaves, not testing advertising. And if there is no evidence that the test scores are predictive of real-world sales effects, the test is of little value.

3. **The Intangibles.** Some researchers are more sensitive to the psychological and political issues in copy testing and will make your life easier. They recognize that testing advertising is as much art as it is science. Results should be an aid to decisions, rather being than the final word.