

# New developments in TV copy testing promise better measures

by William H. Van Pelt

*Editor's note: William H. Van Pelt is senior vice president of Gallup & Robinson, Inc., Princeton, New Jersey.*

**I**n a way, there is nothing new about new developments. They have always been the source of improvements in TV commercial copy testing:

—Back in the 1950s, when commercial testing began, two fundamental approaches evolved. One, forced-exposure, in-theatre testing, was used primarily by those interested in attitude change and in larger sample diagnostics; the other was real-world, on-air testing where the primary measures were intrusion, or recall, and communication among recallers. In those days, most

shows had a single sponsor and on-air testing was usually on a custom basis in a post-test mode.

—During the '60s, the first syndicated service emerged that tested virtually all the commercials aired on prime time. This lowered testing costs and accumulated masses of data that could be analyzed to discover the techniques that promoted or inhibited effective recall. Measures were expanded as Persuasion was added to Recall, and the first validation studies linking the performance measures to sales were completed.

—The '70s gave rise to low-rated independent and cable channels which opened the opportunity for on-air pre-testing through narrowly exposing commercials at low cost. The technique of inviting viewers to watch an on-air program at home successfully raised the average recall score from about 10% to almost 30%, thus allowing more precise measurements, particularly among weaker commercials. Also during this

time, the technique of testing rough commercials on air was introduced, making pre-testing economically practical.

—In the '80s, the techniques themselves remained very much the same. However, we also saw the increased use of customized designs, target group testing, particularly for communication and reaction, and pre-testing.

Today, we may be on the threshold of a generation of copy testing progress made possible by promising new capabilities. Important learning both challenges and confirms conventional thinking about how copy testing should be done. This article focuses on the what and the how of these developments, their implications for further improving copy testing, and why our firm has incorporated them into a new copy testing system.

## Current needs and opportunities for copy testing

The best techniques evolve from the

needs and environment of their times. Recently we held extensive discussions with a variety of major television advertisers and found:

—fundamental agreement that copy testing not only was used, but that it was relied upon and that it would continue to be used in the foreseeable future;

—there was a need to improve test/re-test reliability and to better pinpoint evaluative strengths and weaknesses, as well as to give better diagnostic understanding and creative direction;

—there were fundamental disagreements over which attitudinal measures work best, and how Persuasion should be measured.

Concurrently, the past few years have seen several developments that offer major opportunities. Chief among these might be:

•The development and widespread use of new electronic channels for television viewing, such as cable and now the phenomenal growth of VCR tech-

nology. More than 70% of the households now have VCRs, more than cable, and strong growth is expected to continue.

•The substantial progress made on the theoretical level of how advertising works. Concepts such as the Elaboration Likelihood Model (ELM)<sup>1</sup> and Attitude to Ad (A Ad)<sup>2</sup> introduce significant new tools for reconciling competing views on how advertising contributes to attitude formation.

•And on the empirical level, the recent results of the milestone ARF Copy Research Validity Project, which provide important new learning about the basic value of copy testing itself and the usefulness of its various measures.

Because the ARF study objectively and empirically seeks to resolve the market's fundamental disagreement over which measures work best and are most predictive of sales, a considered review of the study is worthwhile.

#### ARF study examined

The final report of the ARF Copy Research Validity Project was presented in July 1990 at the Copy Research Workshop in New York City. Russell I. Haley, professor emeritus at the University of New Hampshire and principal analyst on the study, presented a recap of the origin of this eight year effort and its design, reviewed the results and highlighted the implications of the findings.

The summary of the objective and method that follows is excerpted from the Executive Research Digest of the ARF.

"The objective of the study was to determine the predictive validity of various types and measures used in copy research. Five products were included. Each submitted pairs of commercials which had demonstrated significantly different levels of sales response in one-year split cable sales tests. These were then tested across six different copy testing methods. The objective was to determine how successfully each of the various copy testing methods and measures predicted sales 'winners.' The copy testing methods included all major types of measures but did not necessarily reflect the proprietary differences in

measures used by specific copy testing firms. Moreover, as Haley noted, 'It is always possible that with other commercials, other brands, other markets or other question phrasing, or in other time periods, different relationships would be found between sales and the measures used in this experiment. On the other hand, the measures that do show strong relationships to sales performance are certainly worth your attention.' Analyses were conducted on differences between on-air versus off-air methods, pre/post versus post only, and single exposure versus re-exposure. All brands were packaged goods and established brand names."

The total sample was approximately 15,000 respondents: 5 pairs of commercials were tested in six copy testing methods, comprising 30 cells of 400-500 interviews each.

The results were both informative and surprising. Below are listed the highest sales predicting measures found in the ARF study.

Measure	Predictive Index*
Likability	300
Recall	234
Positive/negative diagnostics	234
Main Point Communication	188
Persuasion (brand rating, post only)	184

\* An index of 100 indicates that the measure's ability to successfully identify the sales winner (80% confidence) is only operating at chance or random levels. Thus, the chart shows that the measure of Likability was able to predict the winner three times higher than chance.

The study found that a combination of several surrogate measures is more powerful than any one sales predictor alone. For example, Liking and Recall combined could accurately predict the sales winners at a predictive index of 466 (or in 14 out of 15 commercial pairs).

The study leaves most people with two interconnected conclusions:

1. Differences in advertising copy alone can be important, as evidenced by the fact that pairs of commercials were found that under rigorous testing produced large sales differences; and
2. Copy testing works, as evidenced by the fact that a variety of the so-called surrogate measures can predict which

of independently tested pairs of commercials generated incremental sales.

The study also raises a number of significant questions about how people have thought about copy testing, chiefly:

1. The relative strengths of Liking and Recall vis-a-vis persuasion; and
2. The relative weakness of pre-post brand switching vis-a-vis post-only brand rating as measures of persuasive affect. Less apparent issues are the importance of sampling to copy testing validity, and that hand-in-hand with validity we still need understanding to interpret and improve performance.

To hear that several measures are better than one is not surprising. We have been convinced of the relative importance of both Recall and Persuasion particularly in combination, ever since our validation work was completed in the early 1960s.

Our validation procedure compared Change in Advertising Effect, as measured by Recall and Persuasion, versus Change in Sales, as reflected by Last-Time Purchase. The evaluation procedure covered only network prime-time

*Perhaps in this day and age of zapping, muting, and increased clutter we shouldn't be surprised that Commercial Liking may be an increasingly important determinant of viewer attentiveness and message receptivity—and therefore, sales. But there almost certainly remains the danger of form over substance if carried too far.*

television and could not take into account daytime results or other media such as newspapers and magazines. Nor could we factor in such extremely important influences as differences in price or other promotional activity.

And yet, significant positive correlations were found between Change in Advertising Effect and Change in Sales. As for the predictive power of the various ARF measures, many found the relative weakness of the pre-post persuasion measure surprising; we were more surprised by the relative strength

of the Commercial Likability findings. Our use of this kind of questioning has caused us to differentiate between product-based commercial liking and entertainment-based commercial liking. We tend to be cautious when the commercial simply entertains the viewer without making the brand or a brand benefit an integral part of that entertainment. Perhaps in this day and age of zapping, muting, and increased clutter we shouldn't be surprised that Commercial Liking may be an increasingly important determinant of viewer attentiveness and message receptivity—and therefore, sales. But there almost certainly remains the danger of form over substance if carried too far.

Is Likability a better measure of affect than others, including traditional persuasion measures, or is it related to some additional dimension in the communication dynamic? There is a good deal of evidence in the cognitive response literature that suggests that when consumers actively process messages, they are more likely to act on them than when they passively receive messages. One way copy testing has traditionally measured this is by reviewing the respondents' verbatim descriptions of their reactions to determine the extent to which they reveal processing that goes beyond mere recitation of the ad's content.

The subjectivity of the coding process associated with the various systems has, however, presented considerable reliability problems. If the Validity Project's Likability is a closed-end surrogate for this open-ended involvement analysis, it may add new vitality to an important dimension of the communication dynamic.

Our own thoughts on why Liking is important are summarized below:

—Commercials that are liked may get processed more fully. Liking may reflect the degree of positive cognitive processing or viewer involvement that occurred.

—Liking may be a measure of the positive affect that has been transferred from the commercial to the brand as a part of attitude formation. Liking a commercial becomes a salient attribute of the brand and/or evokes a gratitude response, particularly in low-involvement

categories (the so-called "A Ad" phenomenon).

—Commercials that are liked may get better exposure. During second and subsequent exposures, viewers may be less likely to mentally or physically screen out the well-liked commercial and be more willing to watch it again.

—Liking may be a more benign measure of Persuasion than buying intent or brand switching in that the respondent does not feel he or she is being asked to buy, or make some commitment to buy, the advertised product. Liking may be Persuasion in the vernacular of the respondent.

#### Important criticisms

The Validity Project is subject to important criticisms. Certainly the core questions about the replicability of the five-pair test and the applicability of the findings from these specific tests to other product categories and market situations will be debated within many councils. Another debate will surely involve which combination of predictors is most valid for a given product category.

Additionally, we will all try to judge to what extent design differences between the prototype methods used in the project and those that we individually use should be expected to weaken or strengthen the relationships suggested by the findings.

There is, of course, an important difference between Validity and Understanding. Validity in advertising research has to do with whether or not the measures selected are relevant to sales. Understanding has to do with what actually happened during exposure in the sense of being able to interpret and use that information to improve advertising effectiveness. Although we are all farther away from understanding the dynamics of the communication process than we would like to be, efforts like the ARF Validity Project allow us to test and refine our hypotheses about the most basic of all copy testing issues: how advertising works. Properly considered and judiciously applied, the study has valuable implications for all of us involved in advertising research.

#### Subsequent work on Liking

Since the ARF released its findings, we have done additional work to better understand the new measures, particularly Liking. One of the questions about Liking that we have studied is how it relates to the more traditional surrogate measures of advertising effectiveness, Recall and Persuasion. Does Liking provide insight about some new dimension of the communication process or does it measure, perhaps more effectively, one or more of the dimensions currently being considered?

In the pilot work for our new television copy testing service, InTeleTest, we tested eight commercials from six different product categories. We obtained Recall, Liking and Persuasion, as measured by overall Brand Rating. Analyzing the results at the respondent level, we found very strong positive correlations between Liking and Persuasion, and very weak correlations between Recall and either Liking or Persuasion. This confirmed previous work our firm did in the mid-'70s that showed a lack of correlation between Recall and Persuasion.

From additional analysis conducted in 1990 on Gallup & Robinson and ARF measures on print advertising, we found comparable conclusions: a significant positive correlation between Liking and Persuasion, and a relative lack of correlation between either Liking or Persuasion and Recall.

	Television	Print
Liking and Persuasion		
Liking ↔ Brand Rating	+ .66	+ .60
Recall and Liking or Persuasion		
Recall ↔ Liking	+ .13	+ .24
Recall ↔ Brand Rating	+ .10	+ .21

This shows Liking to be strongly related to the affect measures and that it and Recall are measuring different dimensions of ad performance. Further, we find that adjectives associated with positive Liking are generally those that are associated with information rather than the entertainment dimension of the commercial. Liking does, indeed, seem to be an important measure of advertising effectiveness.

### Applying the findings

The confluence of market needs and copy testing developments makes our field as exciting and potentially fertile a research discipline as it has been in years. It also puts research professionals in the position of again having to decide between maintaining the status quo in what they use, particularly for continuity of norms and user acceptance purposes, or experimenting with a new design. As a supplier, our own decisionmaking process has led us to InTeleTest, a new copy testing system which we feel offers at least three improvements over conventional approaches:

1. New and expanded measures. The ARF study shows that the traditional measures of Recall and Persuasion continue to demonstrate empirical validity. It also shows that new and expanded measures add important dimensions to understanding the full communication dynamic of a commercial.

2. Controlled at-home exposure. Up to now, an advertiser had to choose between testing on-air (with the plus of being in the environment in which the stimulus will appear) or testing in a theater/mall (with the plus of being more controlled). The widespread acceptance of the VCR opens the opportunity for a new form of copy testing distribution that combines the real-world advantages of on-air testing with the control advantages of theater/mall testing.

Using a VCR technique, test commercials can be inserted into program material that has never before been seen. It offers a number of benefits and improvements over current on-air, or in-theatre exposure media. Importantly, it still maintains an in-home, natural viewing situation. The cassette, contrary to what we initially thought, does not hype recall or other affect levels.

Unlike on-air testing, however, the cassette approach allows the same program to be used from test to test and the commercial environment to be selected by the researcher, thus giving laboratory control over the exposure while maintaining a real-world setting and thereby improving reliability.

The InTeleTest cassette allows for re-exposure of test commercials so that measurements and reactions can be reported for the total sample and not confined to recallers.

3. Better samples. Sampling is among the most murky issues in copy testing. There are at least two aspects of sampling where improvement can be sought.

First, we should push ourselves away from the severe city limitations that copy research has come to accept. Our past experience, over thousands of commercial tests, shows that performance of the same commercial in different markets can be significantly different. Russ Haley echoed this in his ARF presentation when he said, "It is a well-known fact that copy test results for the same piece of copy can and often do vary from market to market."

We feel InTeleTest testing, in 10 markets dispersed across the U.S. including major metro areas, is a more representative sample than can be offered by on- or off-air services that are restricted to two, three, or four markets.

Second, we need to find ways to boost respondent participation rates in our studies. When mail invitations or telephone pre-recruiting are used, the rates at which people agree to participate are sufficiently low to throw into question the representativeness of those who participate versus those who do not. We have found that personal placement and the invitation to view a never-before-seen pilot show increases acceptance significantly. The result, we feel, is more representative samples.

### Looking ahead

Looking ahead, we can't envision new techniques 20 years out. It seems likely, however, that new measures based on improved understanding of how advertising works, better ways for distributing test commercials, and improved sampling should guide our thinking for years to come. And even newer developments will surely emerge. For in a way, there is nothing new about new developments. □

1 (ELM: Petty and Cacioppo, 1981a)

2 (P. Miniard, S. Bhatia, R. Rose: *Journal of Marketing Research*, August 1990)