JOEL S. DUBOW

POINT OF VIEW: RECALL REVISITED: RECALL REDUX

JOEL S. DUBOW
Assistant Professor of Food Marketing
Saint Joseph's University
and
Marketing Research Associate
The Canaan Parish Group

Day-After-Recall (DAR), as a means of determining the sales effectiveness of television commercials, is reviewed from the perspective of how DAR is measured during the 1990s as opposed to a decade earlier when the (now dead) Burke DAR system was dominant. In addition to questioning the validity of much of the reliability-impugning research which has been reported, it is shown that the variables which threatened the reliability of Burke DAR are better managed by today's surviving systems, such as Gallup and Robinson, Mapes and Ross, ARS, and ASI. Prior validity studies which have been reported as showing "checkered" results are sorted out by Burke versus the other systems; the studies showing no validity are all noted to be Burke-based, while the several which did show validity are seen to come from the non-Burke pool. Reanalysis of prior research, along with the addition of more recent studies, leads to discussion regarding the shape of the recall function and how recall scores ought to be interpreted. Arguments are presented which (1) support a threshold, rather than linear, recall-to-sales function, and (2) speak to the necessary-but-not-sufficient argument for recall. Additional discussion addresses the issue of which measure, among those that DAR can generate, may be the most propitious. Both applied and future research implications of this revisitation of recall are discussed.

Things change! When the *Zeitgeist*, the spirit of the times, changes, we come to view things differently. It is startling to note how different can be the conclusions we derive from the very same data when we view them through a different prism. Such is the case with the concept and measurement of "Recall." This article will examine recall through an early 1990s prism, in contrast to the prism through which Gibson (1983) viewed recall in his notable *JAR* article titled "Not Recall." In a scathing review, Gibson impugned the reliability, validity, and theoretical basis of recall with regard to its usefulness as a technique for assessing the worth of television commercials.

What has changed in the decade since? For one thing, with the passage of time the flames of the debate between recall and alternative measures, such as attitude and persuasion, have died down. In the cooler, quieter atmosphere of today, we can see things more dispassionately, and, I contend, more clearly. As we look back over the recall conflagration of the early '80s we will see that the debate generated more heat than it did light.

We will see that some of Gibson's references might be interpreted quite differently. The peer review process also forces us to pay closer attention to those drearily dull, but epistemologically necessary, concepts which we call underlying as-

Ultimately, the research on reliability and validity which began to emerge in the 1960s and beyond was not of the P&G DAR system—it was of the Burke DAR system.

sumptions and operational definitions. We will see that data which has been used to impugn the reliability of recall has lacked certain characteristics necessary to using product-moment correlation as an assessment measure and that alternative statistics may be called for; we will see that the assumption of a linear relationship in quantifying the validity of recall may have led to an underestimation of the validity of recall; and, we will see that all recall measures are not created operationally equal.

To this latter point, we note that Gibson lumped all recall systems together by using a conceptual definition, while, in fact, critical differences exist in their operational definitions. Today, Burke DAR is dead, and much of the reliability impugning research was Burke-based. The major systems which displaced Burke are operationally different. The Burke system employed a *post*recruited audience and *on-air* exposure; today's survivors use either *pre*recruited audiences, *controlled* exposure, or both. We will see that by failing to attend to these differences, "Not Recall" painted recall with too broad a brush.

Another change in perspective is how we define the copy-testing problem. In 1983, it was still largely "Recall *versus* Persuasion"; today, we find ourselves, benefit of the ARF's Copy Research Validity Project (Haley and Baldinger, 1991), thinking so much more in terms of multiple measures. Our problem definition now operates in a "Recall *and* Persuasion" framework. We will see a most interesting example of how asking the question differently can lead us to a quite different conclusion—from the very same dataset.

A Reliability-Attendant Oral History of Burke Day-After-Recall

The following oral history offers a "What went wrong?" examination of Burke DAR. Oral histories vary with the orator. Perforce, they suffer from limitations imposed by the methodology: peoples' memories (and biases), the interrogatory venues (frequently, cocktail lounges), and the modes of transmitting information (fone, fax, and "Federal"). Therefore, we will limit our oral history to those aspects pertinent to the reliability of Burke DAR, thus eliminating most sources of disagreement among our oral historians: Randy Brooks, formerly of Burke Marketing Research; Bill Greene, recently retired from Gallup and Robinson; and Meg Blair of research systems corporation.

Emergence. What came to be the Burke DAR system emerged in the early 1950s at Procter & Gamble. Gallup and Robinson (G&R) had several years earlier developed the day-after-recall method for print advertising and was working on a television adaptation. Following a G&R presentation, P&G decided to go it alone by developing their own in-house system. (G&R went on to develop their "Total Prime Time" system, a syndicated service that monitored recall of commercials in a tracking-like manner as they appeared in the natural flow of the media plan, and later their "In-View"—for "invited viewing"—System as a pretesting method.)

The early 1950s were a simpler era than today. Three networks and a handful of independents in any given market carried all of the programming. (Do you remember when TV sets carried only channels 2 to 13?) Programming was more homogeneous (and so was marketing—it was still the era of traditional mass marketing). The entire family sat together watching television on the one/only set they owned. Network shows were "sponsored"—a single company sponsoring the entire show and running only its own commercials.

Advertising was simpler, too. Commercials of the era attended largely to rational appeals, employing product attribute or benefit appeals which could be nested into commercials rather easily and which required rather simple interrogation of consumers in order to assess whether they "got the message." Indeed, playing back the message was tantamount to "getting" the message. In the context of how P&G viewed the task of advertising—to get a key message about the product into consumers' minds—the system certainly had content validity. But, due to P&G's well-known penchant for secrecy, we do not know whether it had predictive validity for sales. Our historians believe it did.

Evolution and Growing Pains. Yet, as studies began to emerge from other sources some 15 to 20 years later, there was reason to question the validity of the recall measure. To understand why, we must note (1) that those studies were not from P&G, and (2) the system, as practiced external to P&G, had by that time evolved into a less consistent one with regard to test execution and a host of derivative variables.

Here's how that evolution occurred. Not long after P&G developed the system, a business decision was reached to divest of the in-house DAR fieldwork. A P&G employee, Alberta Burke, was sanctioned to establish a company which would conduct P&G's interviewing. Enter Burke Marketing Research. Contemporaneously, P&Gers who had left the fold to join other companies found themselves in need of advertising research and went to Burke as a supplier. The business grew over time—and growing pains emerged. Consider: P&G's system was a single-city, single-interview location, constant program type, closely exposure-controlled system; but, as Burke's business grew, these constancies dissolved. More cities were added to accommodate the demand, and additional interviewing services, less closely controllable and less consistently trained, were added. Different shows were used by different companies, adding show type and concomitant audience differences as variables. There was less control over commercial-exposure conditions as sponsorships became shared. Ultimately, the research on reliability and validity which began to emerge in the 1960s and beyond was not of the P&G DAR system—it was of the Burke DAR system.

The Critics. And, who was responsible for much of the early research? Not the advertisers, but their agencies—who suffered from and resented being "Burked." [burke; *v.tr.*: to murder, esp. by smothering.] In addition to being motivated to burke Burke, the agencies had the means. In contrast to the advertisers, who had only their own limited DAR databases, the agencies had collected large bodies of trans-client data upon which they could conduct their analyses.

The Competitors. Burke came to find itself under siege, but was slow to react. An opportunity emerged for other companies, knowledgeable of the nature of Burke's problems, to develop competitive DAR systems which avoided those problems. Among these companies/systems, four notable ones that survived Burke, were (1) Gallup and Robinson's In-View system, along with (2) its near-clone from Mapes and Ross (M&R), (3) ASI's cable-based system, and (4) research systems corporation's theatre-based ARS system. Stewart, Furse, and Kozak (1983) have documented the characteristics of these several systems, and we can examine them with regard to how they are better insulated than was Burke DAR against a variety of exposure, respondent, and measurement variables.

G&R's In-View system and the Mapes and Ross system differed from Burke DAR in three key ways: (1) both employ *prerecruited* audiences to watch the designated show, (2) both use a relatively constant program format from test to test, and (3) both place the advertising themselves. G&R has mostly been a one-city service (usually the same city), while M&R runs a three-city test (usually in the same three cities).

ASI exerts even more control by prerecruiting its sample to watch a prepackaged show placed on an unused cable channel. The show is constant from test to test, thus also eliminating any variability due to programming. Variations in exposure conditions are minimized by rotating the commercials' exposure conditions across the two cities used in the test. The cities are selected from a small pool of cities and may vary from test to test.

ARS differs from the other systems by virtue of exposing their prerecruited audience to the advertising in a theatre situation. The commercials are nested into a constant program, and exposure conditions are rotated across a set of four cities in order to balance-out variations in exposure conditions.

By virtue of using prerecruited audiences, all four of these systems do away with virtually all of the test-to-test variation in audience composition which plagued the Burke system. G&R and M&R, by virtue of self-placing the advertising, *minimize* several of the exposure variables, while ASI and ARS actually *control* them. All four, by virtue of being syndicated systems, as opposed to Burke's custom service, also effect more test-to-test control over their interviewing and measurement procedures.

Each of these systems was, indeed, operationally quite different than Burke DAR, and there was available evidence prior to "Not Recall" to indicate that such differences could have a rather large impact on the reliability of recall measures. The evidence comes from Achenbaum, Haley, and Gatty (1967), who demonstrated that on-air exposure situations would require sample sizes four times as large as those used in forced-exposure systems in order to obtain the same levels of reliability. By not being attentive to such differences, "Not Recall" painted recall with too broad a brush.

We will now return to examine the same research Gibson reviewed in "Not Recall," but through a 1990s prism. (Some might choose to say "with 20/20 hindsight.") Though Gibson criticized DAR on three grounds, reliability, theory, and validity, we will dispense with debating his arguments regarding the theory of recall for two reasons: (1) the theory which Gibson disputed was Gibson's own interpretation of other people's theories, and (2) it was largely based on critiquing the logic of measuring recall and persuasion in the same system—but that argument dissolves if you measure them independently, as your current author did in his work at Coca-Cola (Dubow and Stout, 1983; Dubow, 1984, 1986).

Reliability Revisited

In reexamining the reliability citations in "Not Recall," we will exercise a good bit of quality control. We will address the fact that many of the citations were not evidence-based, but have tended to take on an evidential appearance by virtue of secondary referencing—some over and over again. We will note one interesting case of good research (Clancy and Kweskin, 1971) which has been often miscited. We will debate the conclusions reached in another oft-cited work (Yuspeh, 1979). And, we will examine the concept of using the correlation coefficient as the preferred means of assessing reliability; in the later regard, we will come to different conclusions regarding reliability than "Not Recall," derived from the work of Young (1972) and Clancy and Ostlund (1976).

There were two thrusts to "Not Recall's" attack on the reliability of recall. One was based on presenting an inventory of all the research which identified variables that were shown to adversely affect recall's reliability. The second thrust was to impugn the nature of the high reliability coefficients which had been reported for Burke DAR. In addressing each thrust we will partially agree and partially disagree with Gibson.

Reliability-threatening Variables. "Not Recall" treats us to an expansive three-part table (Tables 5, 6, and 7 in the original article) which cites a dozen sources, addressing 19 exposure/respondent/measurement variables which can impact recall scores and vitiate their reliability. Were even half of the resultant 42 source × variable entries to survive close scrutiny, the argument would be quite convincing. And more than half (actually 24) do survive close scrutiny. Most convincing is that 18 of the surviving entries come from DAR vendors (Burke, 1975, 1976, 1980; ARS, 1979, 1980). However, there are two problems with the set of 42 entries. The first is that they impugn only the Burke DAR system; as suggested above, Burke's surviving competitors better control them—which is probably why they survived. The second problem attends to a lack of either substance or validity to many of the 42 inclusions—13 lack substance, 4 come from substantive research but are erroneous, and 1 is all but invalid.

Fully 13 of the subject citations lack substance by virtue of being non-evidence-based "mentions" of variables, apparently from the speakers' and authors' experiences, offered sans data. Though there is no reason to question the truth of these mentionings, the problem that occurs is that they have come, apparently due to continued secondary referencing, to be passed on in the subsequent literature as if they were evidential. For instance, Heller (1971), to whom Gibson assigned 6 of the 42 entries, likely never expected to be significantly referenced. He described his offering as a "talk"—delivered to a luncheon meeting of the Advertising Effectiveness Group of the New York chapter of the American Marketing Association. In that context, he was speaking in part to share his experiences and in part to entertain. And, in that context, he did, indeed, mention six variables which in his experience appeared to impact recall scores—but he presented no data. In a similar vein, the two Lamar (1981) inclusions, which occurred at a conference of the Association for Consumer Research, were merely parenthetical comments, also offered sans data.

Other non-evidence-based inclusions among the 42, however, come from journal articles. Three of the four attributed to Young (1972)—position in program, time of exposure, and age—were stated without evidence; how-

Exhibit 1: Reproduced from Clancy and Kweskin (1971) Multiple Regression Analysis

Recall	Beta	Partial r	*P* Value
% Rating program "favorite"	.47	.42	<.05
Mean education	−.29	−.30	>.05
Mean age	.05	.06	>.05
% Test brand users	.05	−.07	>.05
% Viewed entire program	.01	.02	>.05
Multiple R = .81			
Multipe R^2 = .65			<.01

ever, the fourth, city differences, was quite well supported with evidence. The two inclusions from Clancy and Ostlund (1976) were also nonevidential.

One reference, given five inclusions in "Not Recall," is evidence-based, but four of the inclusions are wrong. This is Clancy and Kweskin (1971), whose study gives us insight into how misinformation may be created and disseminated. The error is not Clancy and Kweskin's, who did, indeed, investigate five variables—but only one, program attitude, was found to be significant. Their actual table is reproduced here as Exhibit 1, and they concluded thusly: "The results of this analysis demonstrates that at least *one* [emphasis added] factor, program attitude, is significantly related to on-air recall scores."

The paper trail is worth following. The earliest miscitation seems to be by Young (1972) in her so-frequently-ever-since-cited *JAR* article, "Copy Testing without Magic Numbers," wherein she cited Clancy and Kweskin as reporting ". . . that recall scores are subject to *five* [emphasis added] other external influences: program liking, education, age, test brand usage, and viewing of the entire program." (We note, for the record, that in the very same *JAR* issue, Schulman [1972] did cite Clancy and Kweskin correctly.)

Perhaps the most unique miscitation of Clancy and Kweskin was that by Clancy, himself, in Clancy and Ostlund (1976): "Evidence that audience composition can have a profound impact on on-air test scores has been provided by Clancy and Kweskin (1971) and Young (1972)." In addition to noting that his earlier two audience-composition variables, age and education, were not significant, we might also note what Clancy saw in the Young article that led to the use of the term "profound." It appears to be the following: "A second *unpublished* [emphasis added] examination of recall in many tests concluded that recall scores are also subject to systematic biases from three external sources . . ." But Young provided neither a reference nor data for this study which Clancy took as providing evidence of a "profound" impact.

Exhibit 2: From Yuspeh (1979)

☒ = Commercial Scores Affected by Specific Shows

	Brand recall	Play back	Buying intent	Brand Perceptions		
				A	B	C
Sponsor #1 (Males)	X	X	X	X	X	X
Sponser #2 (Males)	X	X		X	X	
Sponser #3 (Females)	X	X	X	X	X	
Sponsor #4 (Females)	X	X	X	X		X
Sponsor #5 (Males and Females)	X	X	X	X	X	
Sponsor #6 (Males and Females)	X	X				X

We come now to a final, very evidentiary, and since oft-cited study. Though procedurally well-designed, this study suffers from a rather misleading analysis and, thereby, a misleading conclusion. We refer to Yuspeh's (1979), "The Medium Versus the Message: The Effects of Program Environment on the Performance of Commercials," a report of a study conducted in 1977 by J. Walter Thompson and six (unnamed) leading advertisers. At issue in this article was not the impact of *program type* on recall, but that of *specific shows within a program type.* If significant, this finding would impugn even the on-air, prerecruited systems (G&R, M&R) which attempted to minimize the potential impact of program type by using a constant program type, albeit with the specific shows within type left to vary.

We first look at what Yuspeh concluded and what was shown in support of the conclusion. Yuspeh claimed the study provided "powerful evidence" to the effect that ". . . the *same* commercial can get radically different recall from different shows even when those shows are the *same type* of show." To reach this conclusion the study examined six measures across six sponsors, producing 36 opportunities to demonstrate significant differences. In Exhibit 2 we reproduce Yuspeh's final summarizing graphic, about which she stated:

> . . . the weight of this evidence makes it abundantly clear that on-air scores are . . . affected by the specific show as well as by the commercial being tested.

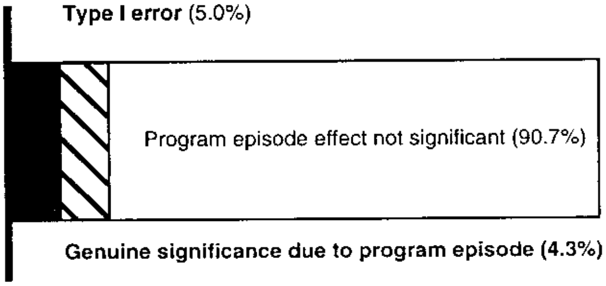
Now, 28 significant hits in 36 opportunities certainly looks impressive. However, there is less here than meets the eye. You see, each box shown in Exhibit 2 was actually a product of *six replications* of the 6 × 6 grid. In addition, Yuspeh utilized the 80 percent confidence level for each of the replications—and, if *any one* of the six replications was significant (not an average, but any one), the box was checked). Now, consider this: The chance of *not* getting at least one random hit in six attempts when using the 80 percent confidence interval is $(.8)^6$, or .26; therefore, the chance of *getting* at least one hit is $1 - .26 = .74$. Thus, the most likely *random* outcome for the 36 box grid was $.74 \times 36 = 27$ hits. Yuspeh's actual count was 28!

A more proper analysis would have attended to the results for all (6 × 6 × 6 =) 216 boxes across the six replications. Because those data are available within Yuspeh's article, we can do that analysis—and we reach a quite different conclusion. Our null hypothesis is that by using the 80 percent confidence interval we would by chance alone get 20 percent significant hits; our alternative hypothesis would be to expect significantly more than 20 percent such hits. In fact, the data show that 23.6 percent (51 of 216) hits occurred. A one-tailed *t*-test tells us that the difference between 23.6 percent and 20.0 percent is not statistically significant: $t = 1.245$, df = 215, $p > .10$.

However, we must note that there is *statistical* significance when we apply the 95 percent confidence interval to the 216 box grid—but it occurs at such a minor level to be *not managerially significant*. Yuspeh found 9.3 percent hits (20 of 216) using the 95 percent confidence interval, and that is significantly greater than the 5 percent that would be expected by chance: $t = 4.677$, df = 215, $p < .001$. But to understand the actual worth of this difference, we need to examine it in the context of all 216 tests. We do so in Figure 1, where the solid portion of the bar represents the 5 percent chance hits, the cross-hatched section represents the 4.3 percent significantly above chance hits, and the remainder represents the 90.7 percent of the time that no significance was obtained.

Figure 1

Yuspeh Data Reanalyzed at .05 Level



It is hoped that future researchers and reviewers will take note of the above discussion of non-evidence based, miscited, and invalid references which have been continually passed on—and will refrain from continuing to cite them in that manner in the future.

Reliability Coefficients. The second thrust of "Not Recall's" attack on the reliability of recall came in the form of questioning the value and meaning of high reliability coefficients reported for recall. Gibson used two examples to show that, examined in more depth, the reliability coefficients may have actually delivered less than they appeared to promise. The two examples came from the aforementioned articles by Young and by Clancy and Ostlund.

Before revisiting those studies we need to make a detour and visit Silk's (1977) *Journal of Marketing Research* article titled "Test-Retest Correlations and the Reliability of Copy Testing." Though written in the pedantic style common to *JMR*, the article is worth the reading in order to appreciate how an overreliance on product-moment correlation can mislead us.

Silk advises us that "correlations often are reported as measures of the reliability of copy testing procedures with little or no attention to the conditions under which such an interpretation is meaningful." He points to the "commonplace practice of using *ad hoc* collections of heterogeneous test-retest data as a basis for computing correlations." His discussion begins by addressing the difference between test-retest reliability as formulated in the psychometric literature and as practiced in the domain of psychological measurement, in contrast to what may happen from analyzing *ad hoc* collections as occurs in copy-testing research. To this point, he particularly points to violations of the equivalence of test-retest conditions. One such frequent violation has to do with the time lapse and, thereby, repeat-exposure opportunities, which occur in many of these data sets.

Silk also addresses problems related to the form of the data sets and the impact of violating the underlying assumptions of correlation. If data sets lack homoscedasticity, they can, depending on the manner in which the deviation occurs, either over- or under-estimate the true reliability (meaning reproducibility) of a measure.

What Silk offered is more easily depicted visually than explained through algebraic proofs. In Figure 2 we present a scattergram for a set of 36 "test-retest" pairs, but we depict them

It is hoped that future researchers and reviewers will take note of the . . . discussion of non-evidence-based, miscited, and invalid references which have been continually passed on—and will refrain from continuing to cite them in that manner in the future.

in a manner in which we can also select out two subsets of 18 pairs each—the solid and the unfilled circles. The unfilled circles represent a "distribution of the extremes," while the solid circles represent an "attenuated distribution." In Table 1 we present the key statistics for each of the three distributions. Notice that all three distributions have exactly the same regression line and virtually the same standard error of estimate. However, despite the functional equivalence on these two most pragmatic statistics (one used for predictability, the other for estimating error risk), the correlation coefficients and variance explained differ. The "distribution of the extremes" exaggerates the correlation for the underlying total distribution, while the "attenuated distribution" understates it.

In revisiting "Not Recall's" two reliability coefficient examples, we will see one of each way in which the correlation coefficients can be misleading.

Young (1972). In Exhibit 3 we reproduce the data from Young as displayed in "Not Recall." What we see is a "distribution of the extremes." Only one of the ten initial test scores is "average"; in a normally distributed population, given Young's apparent use of plus or minus one sigma, six or seven would be "average." Gibson used these data to emphasize the lack of reproducibility of the scores even though the data came from a data set having a test-retest correlation of $r = .87$. But, based upon Silk and Figure 2, we might surmise that the underlying population from which Young's 10 test-retests came must have had a lower correlation than .87.

While we do not know how Young came to choose these ten test-retest pairs from the apparently larger set she had available, we do know that selection from the extremes (nine of the ten) produces a specious analysis. Based on the phenomenon of regression to the mean, whereby extreme scores tend to retest closer to the mean, Young's data set had a built-in predisposition to find significant differences. In fact, seven of the nine non-average test scores did, upon retest, move in the direction of the cited norms. While Young did have the right idea—testing for reproducibility via t-tests—she needed to practice the endeavor on a more representative data set.

Clancy and Ostlund (1976). Question: Why is it that when you lose something, you always find it in "the last place you look"? Answer: Because once you find it, you stop looking. So it is with Clancy and Ostlund, who showed that test-retest reliability coefficients for recall decrease when the data sets are examined *within* product categories. The supporting data, as displayed in "Not Recall," are reproduced in Exhibit 4. There is nothing wrong with the data. What is wrong is that after having found what they were looking for, a recall-impugning outcome, they stopped looking. Had they further scrutinized what they had done, they might have noticed something else that occurred: an attenuated distribution.

If the attenuation occurred due to the creative restrictions which occur within product categories, recall was being blamed for a phenomenon beyond its control. In order to avoid that, Clancy and Ostlund should have

Figure 2

Scatter Diagram of "Three" Distributions

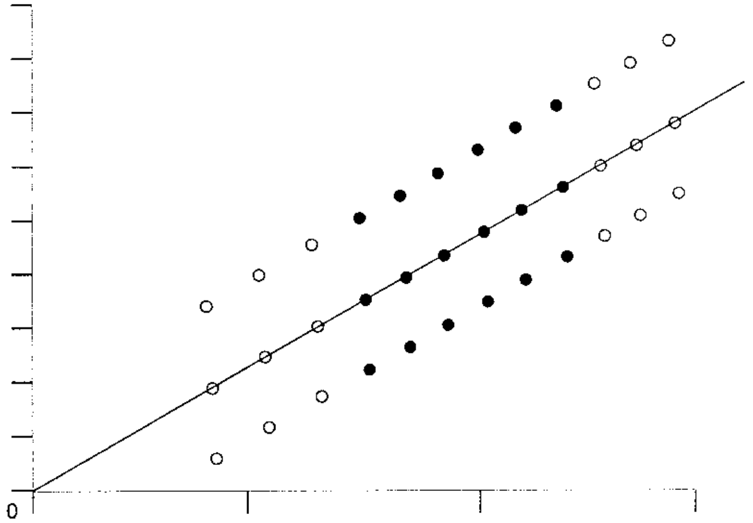


Table 1
Effect of Distribution Violations on Correlation Statistics

Statistic	Full Distribution	Distribution of extremes	Attenuated distribution
r	.82	.88	.57
r^2	.67	.78	.33
$Y' = a + bX$	$0.00 + 1.00X$	$0.00 + 1.00X$	$0.00 + 1.00X$
$s_{y \cdot x}$	2.48	2.52	2.52

done what Young did—conduct a series of t-tests across the full sample of 138 test-retest pairs which they had at their disposal.

To be sure, there is a debate to be joined here: that of DAR system norms versus category norms. We will defer that debate until later, because there is additional data to come which bears upon it.

Current Data. A logical next question to ask is: How does the current reliability data, in the post-Burke era, look when examined in terms of test-retest reproducibility? The two DAR vendors which have thrived since the demise of Burke, ASI and research systems corporation/ARS, have both taken this reproducibility approach in research/promotional materials.

ARS has been doing so for more than 15 years. In the latest dataset released (ARS, 1992), they have reported on 74 cases, comparing the observed distribution of test-retest differences to what would be expected by change. The data were analyzed by both Chi-Square and an F-ratio and found to be nonsignificant. Of the 74 individual t-tests, 9 were significant at $p < .10$, compared to a chance expectation of 7.4.

ASI (1991) has offered data for 31 test-retest cases. They analyzed the data via a Chi-Square Goodness-of-Fit procedure and found no significant difference in the actual distribution of t-scores compared to the expected distribution. They reported that 4 of the individual t-tests were significant at $p < .10$ compared to a chance expectation of 3.1.

Exhibit 3: Young (1972) Data as Displayed in "Not Recall" Recall Data—Test/Retest Reliability

		Scores		Evaluation		
Commercial	Norm	Test	Retest	Test	Retest	
1	37	37	48	Average	Superior	
2	35	46	37	Superior	Average	
3	35	39	30	Superior	Poor	—Different
4	35	25	32	Poor	Average	
5	24	18	24	Poor	Average	
6	24	9	13	Poor	Poor	
7	14	4	8	Poor	Poor	
8	24	29	32	Superior	Superior	—Same
9	24	32	29	Superior	Superior	
10	24	36	36	Superior	Superior	

Recall Reliable! We see now, that by virtue of misinterpretations and obsolescence (because Burke DAR is dead), the pre 1990s reliability-impugning research directed at recall must be relegated to the archives of advertising research. Recall management, as practiced today, is reliable.

However, we need to keep in mind this comment from Stewart, Furse, and Kozak (1983): "Since services are wise to present whatever makes them look good, it is extremely difficult to make meaningful comparisons of alternative services on the basis of the literature typically provided to prospective users." It would not be unreasonable for our industry to set the decision criteria, rather than leave the debate to the competing services.

Validity Revisited

We now turn to the validity of DAR. This discussion of validity, as was (most of) Gibson's, is confined to the relationship of television DAR to sales for existing, not new, products.

"Not Recall" is at its most accurate when Gibson states: "Several researchers have suggested some modest relationship between recall and sales or some other sales-related factor," and "In contrast several researchers have found no relationship. . . ." He used the sum of these studies, running from zero to modest, as evidence of recall's invalidity: "We know that the evidence for the validity of recall is—to be charitable—'checkered'."

. . . by virtue of misinterpretations and obsolescence (because Burke DAR is dead), the pre 1990s reliability-impugning research directed at recall must be relegated to the archives of advertising research.

But Gibson, inattentive to the differences in the operational definitions of recall, failed to notice this interesting circumstance: the zero sources were all 1970 or earlier and Burke-infested, while all five modest sources were non-Burke. They were one from TeleResearch (1971), one from ARS (1980), one from Mapes and Ross (Lamar, 1981), and two dealing with Gallup and Robinson data (Greene, 1981; Stout, 1981). So, for non-Burke measures of recall the data are solidly, well, modest. (Note: the ARS study was for new products.)

That Curious Mapes and Ross Study. The Mapes and Ross study is a most curious one. It has been reported in two venues (Lamar, 1981; Ross, 1982) with different conclusions. After citing the positive statement by Lamar and putting it in the "modest" set, "Not Recall" then used Ross to retract the statement. Two years later, Stewart, Pechman et al. (1985) also mentioned Lamar, reporting him as a secondary reference from "Not Recall," but without discussion, and then proceeding to discuss Ross at length—also accepting his negative conclusion about the validity of recall.

The problem is that the Lamar paper, presented to the 1981 Annual Conference of the Association for Consumer Research, was not included in the proceedings of that conference. Against all odds, your present author was able to obtain Lamar's speaking notes from the archives of the company which underwrote the Mapes and Ross study. (Lamar, himself, is now two companies removed.) The statements and data contained in Lamar's notes, when juxtapositioned against how Gibson cited Lamar, and how Ross analyzed the data, lead me to conclude that the Mapes and Ross study offers more than modest support for the validity of recall—just not as much as it offers for persuasion. In addition, the Lamar archives allow us to tease a bit more information out of the study that has important implications with regard to the manner in which recall interacts with persuasion. (A copy of the Lamar materials has been provided to the editor of this journal in order to verify what will follow.)

We will first deal with how "Not Recall" handled Lamar and Ross and then delve into the Ross study as reported by Ross. Here is what appeared in "Not Recall":

> Lamar . . . reported that recall in addition to persuasion added a "modest increment" to the measurement of commercial effectiveness. Reporting on the same research, however, Ross said, "Proven recall . . . is a very poor measure of a commercial's effect on consumer purchase."

Lamar's notes, however, suggest that he took a stronger position on recall:

> Recall also has a direct relationship to purchase behavior independent of persuasion. However, the persuasion measure is much stronger.

Notice that Lamar used the term "independent of persuasion," not "in addition to persuasion," and that the term "modest increment" appears nowhere in his notes.

Exhibit 4: Clancy and Ostlund (1976) Data as Displayed in "Not Recall"

Recall Data Test/Retest Reliability

	n	*r*	r^2
Data Base 1 (Ad Agency)			
All products	106	.67	.45
Auto products	16	.45	.20
Proprietary drugs	15	.52	.27
Soaps and cleaners	10	.69	.48
Toiletries	15	.81	.66
Average (wtd)	—	.29	.08
Data Base 2 (Research Firm)			
All products	32	.76	.58
Proprietary drugs	7	.57	.32
Toiletries	13	.83	.69
Soap	6	.12	.01
Average (wtd)	—	.59	.35

Ross did not see the data in the same manner as Lamar. In the *JAR* article which reported on the research, Ross concluded thusly:

> The results of this validation study support the hypothesis that changes in brand preference . . . do relate to substantial changes in test-brand buying. However, proven recall emerges as a copy-testing measure that did not significantly discriminate on purchase.

We appear to have a quality control problem here. You see, though Ross asserted no significance for recall, and implied hypothesis testing for persuasion, not a single significance test appeared in the entire paper, and some of the sample-size information necessary to do our own analyses was absent. It appears that Ross' problem in coping with the data came from the way he defined the problem: Recall *versus* Persuasion. As Lamar saw the data, and so do I, they better support: Recall *and* Persuasion.

Let's look closer. Here's how Ross, having adopted a "versus" posture, interpreted the recall data:

> Looking at the results in terms of proven recall performance, proven recallers exhibited somewhat greater test-brand buying levels than those who could not prove recall (38.8 percent versus 25.2 percent). However, this higher buying level can be attributed in large measure to those within the proven recall group who also changed their preference toward the test brand.

There are several problems with this statement. First, there is a cart-before-the-horse aspect to what Ross says. If you want to sing the praises of a *recall-based* persuasion measure, certainly some credit must go to recall. How can you have recall-based persuasion without recall? Second, Ross' use of the term "somewhat" is debatable; a 38.8 percent buying level represents 54 percent more penetration than a 25.2 percent level. And, for the sample sizes in play (n = 807 and 2568), t = 7.09, and $p <$ .001.

Third, Ross ignored other data (Table 4 in his paper) which shows that even when persuasion is held constant, higher buying rates occurred in association with proven recall. Ross' data, with the addition of n's (from Lamar's archival notes), and the differences, t- and p-values which support this statement, appear, herein, as Exhibit 5. For the two situations in which no persuasion occurred ("Did Not Change" and "Preferred Test Brand Pre-Exposure"), recall alone is associated with highly significant differences in brand penetration among category buyers—and where a persuasion change did occur, it is nearly significant.

Exhibit 5

Ross (1982) Test Brand Shares by Recall and Persuasion Groups (*n*'s)

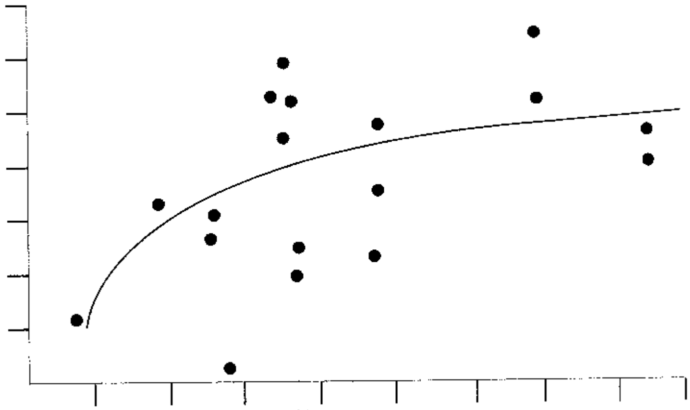
Persuasion status	Proved recall	Did not prove recall	Difference	t	p
Preferred test brand pre-exposure	74.0% (135)	58.6% (716)	+13.4	3.19	<.001
Changed to test brand	45.8% (338)	40.7% (466)	+5.1	1.46	<.10
Did not change to test brand	17.0% (329)	11.7% (1391)	+5.3	2.41	<.01

We leave some additional data from Lamar/Ross for later.

Erroneous Function Assumptions? Let us also address "Not Recall's" characterization of the strength of the relationship be-

Figure 3

Market Share Regression Residuals as a Function of Recall Scores



tween recall and sales as being "modest." This may be numerically true, but the characterization as "modest" may rest on two erroneous assumptions. The first is the either-or assumption, as if one copy-test measure should fully predict the sales function of advertising. In a multivariate, multiple-measures world we would hardly expect any one variable to account for most of the variance. What we ought to expect is several "modest" relationships which *collectively* come to account for a majority of the variance.

In a multivariate, multiple-measures world we would hardly expect any one variable to account for most of the variance.

We may also underestimate recall numerically by using linear correlation to measure its strength. Linear correlation assumes that the underlying function is, well, linear. If that assumption is wrong, the linear correlation coefficient will underestimate the strength of the relationship.

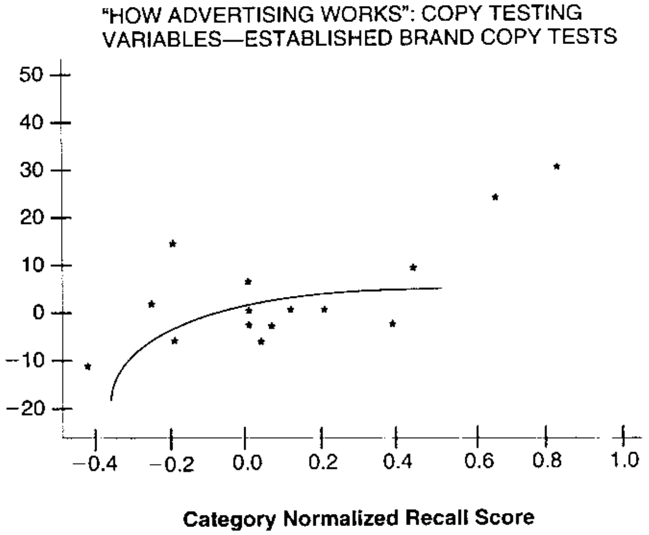
New Data. Let's examine the question of a linear relationship. "Not Recall" reported Stout (1981) as one of the positive sources for recall validity when recall is used in combination with measures of persuasion, advertising weight, and promotional activity. A few years later, having been the analyst for what Stout reported, I reported (granted, as a "mention," sans data) that when these same G&R recall scores are examined alone, the relationship appeared to be curvilinear (Dubow, 1986). The "share" data used to develop the relationship occurred as regression residuals after the effects of advertising weight and promotional activity had been partialled out. We now make the (coded) data available in scatterplot form (see Figure 3). The linear correlation for this array is $r = .53$ and $r^2 = .28$—not very much different from G&R's own earlier validation where $r = .45$ (Greene, 1981). However, if one applies a curvilinear regression line, higher values occur. The curvilinear line which appears in Figure 3 is one of several which were examined—all of which produced analogous values for r in excess of .60 and for r^2 in excess of .36.

This curvilinear relationship, with a very flat function over most of its range, is also supported by IRI's "How Advertising Works" study (1991). The portion of their distribution which appears to cover the same range of scores operant in Figure 3 (recall scores which index below 160 against the norm) permits a very similar curve to be fitted to the IRI data, as shown in Exhibit 6. The IRI data also suggests that scores indexing above 160 may initiate a second underlying recall function—which I have previously labelled the "Twilight Zone" of recall (Dubow, 1992).

A Threshold Function? The steep low-end dropoff for sales in Figure 3 may lead us to suspect that there might be some point at which copy has no impact at all despite weighing in with recall scores above zero. That is, the modus operandus of recall at the low end of the curve would be a threshold function. This suspicion led me to review the company's split-cable testing history. As a matter of practice, when designing those tests, we usually attempted to include a no-advertising panel, allowing us to use many of the tests to answer the question, "Do these advertising arrangements we are testing *have any effect at all?*" In examining the results from these "any effect at all" assessments, there occurred about a dozen cases representing the flat portion of Figure 3—recall scores which indexed above 80—and most of them showed a positive impact on sales.

Only one of those tests used copy which indexed below 80 (specifically, 73) on recall. What were the results? Despite having scored above average on the

Exhibit 6: From IRI (1991)



If you think of recall as the ignition system of an automobile and persuasion as the fuel system, it is clear that the fuel system delivers much more power to the vehicle's function. But the fuel system doesn't get a chance to perform its function if the ignition system is too weak to start the car. Recall ignites persuasion!

company's persuasion measure, there was no significant sales impact over the no-advertising condition. And, the effect replicated, because the copy was tested at two weight levels, normal and double weight for the test brand; neither level produced a sales impact.

The concept of a low-end threshold function for recall brings a qualitative argument to bear on the "modest" numeric validity of recall. If you think of recall as the ignition system of an automobile and persuasion as the fuel system, it is clear that the fuel system delivers much more power to the vehicle's function. But the fuel system doesn't get a chance to perform its function if the ignition system is too weak to start the car. Recall ignites persuasion!

What about the CRVP? A potential bugbear in the recall validity equation is the ARF's Copy Research Validity Project (Haley and Baldinger, 1991). Proven recall, the traditional DAR measure, did not fare well in the CRVP. There is, however, one problem with the CRVP as it relates to our examination of the recall function. Due to the anonymity and data security promised the cooperating companies in the CRVP, the raw data of the study remain unavailable for inspection. If all of the CRVP recall scores fell above an 80 index, they would, in fact, support the curvilinear relationship by virtue of validating the flat portion of the function. (Because the commercials which did enter into the CRVP had passed the several companies' internal reviews for test-market purposes, it is quite likely that they exceeded minimal intrusiveness levels.)

Depth of Recall. The CRVP also contains an intriguing bit of data on recall having to do with the *depth* of recall. By "depth" of recall I mean the extent to which respondents are required to provide proof of recall. Aided brand recall, requiring merely a "yes" or "no" answer to a brand cue, would be less deep; offering a detailed description of the commercial would be quite deep.

The DAR questioning sequence takes respondents through a sequence of cues, with proven recall credited only after maximal probing in order to pull out more depth of recall. But the CRVP found that a simpler measure, "brand recall from a product category cue," was a discriminator of sales effects. Haley and Baldinger assessed the finding this way:

> The measures that perform best use minimal cues. It appears that extensive probing to enlarge the recall base is not a good idea. While probing does provide more people who can be classified as recallers . . . , it does so at the risk of diluting the predictive power of the recall measure.

Table 2
Lamar (1981)/Ross (1982) Persuasion Rates by Depth of Recall

No claimed recall	Claimed, not proven	Proven recall
15.1%	44.5%	49.9%
[n = 1255]	[n = 632]	[n = 677]

Statistics (one-tailed) for adjacent pairs:
t = 132.43 1.96
p < .0001 <.02

There are several additional pieces of research which speak to the depth of recall issue. One is the aforementioned Lamar/Ross research, which contains data that lends support to the "minimal cues" hypothesis. Unreported by both Lamar and Ross, but discernable from Lamar's documents, is the association between depth (not amount) of recall and persuasion. These additional data, displayed in Table 2, show the persuasion percentages for "No Claimed Recall," "Claimed, Not Proven Recall," and "Proven Recall." Though both increments over "No Claimed Recall" are associated with significant increases in persuasion, most of the difference in persuasion comes in moving from "No Claimed Recall" to "Claimed, Not Proven Recall." Moving further to "Proven Recall" adds just a little more.

A second supportive piece of research which suggests that higher cuing may not be productive comes from a study which was reported in "Not Recall" but not cited earlier because Gibson presented it in his "Theory" section. This is General Mills data, reported by Baumwoll (1978), which demonstrates that recall "playback" distributions conform poorly to independently measured changes in brand perceptions which follow exposure to advertising. (I can second this phenomenon from my own experience at Coca-Cola.) This phenomenon does not, however, impugn the validity of recall; what it does is inform about the way we ought to use recall results: Do not rely on recall as a *communications* measure; the utility of recall scores is limited to assessing the *intrusiveness* function.

A third study which addresses the depth of recall issue came post-"Not Recall." This is Walker and von Gonten's (1989) work which demonstrates that a memory trace may occur in recall testing even though respondents cannot prove recall. Most importantly, not all commercials are equal in this regard—meaning that the attempt to elicit proven recall by requiring respondents to verbalize what they remember may add as much "noise" as information to the proven-recall measure.

Conclusions

After closely examining the literature cited in "Not Recall," and benefiting from a more recent perspective on copy testing, along with the addition of more recent data, we find the conclusions reached in "Not Recall" in want of revision.

Our main conclusion is that Day-after-Recall of television commercials, as practiced by today's vendors of recall, is a valid measure of the ability of commercials to impact sales, but it does not bear a linear relationship to sales. The value of DAR is to cull out advertising that has a high likelihood of being ineffective due to its lack of intrusiveness. Epigrammatically speaking, we must replace the term "Not Recall" with the term "Recall Redux"—recall leads the way, after which other, probably more powerful, factors take over.

Whether, or how much, it improves advertising effectiveness to increase recall scores which are already above threshold level remains to be determined. Though the evidence is not favorable, the prejudice shown by earlier juries may warrant a retrial.

> *. . . we must replace the term "Not Recall" with the term "Recall Redux"—recall leads the way, after which other, probably more powerful, factors take over.*

Implications

There are a number of research implications and applied implications which follow from, or are suggested by, our revisitation of recall.

Research Implications. *The Shape of the Recall Function.* The case made for the validity of the recall function is stronger than the case made for the shape of the function. The reason resides in the circumstance that there is so little data available for low recall scores. It would appear that advertisers have, in fact, been using recall to cull out highly probable losers. In order to verify the low-end curvilinear/threshold function we need to amass more data for low DAR-scoring commercials. But that may be unrealistic; we would hardly expect advertisers to place on air putatively weak advertising solely for research purposes.

The higher end of the recall function may be more tractable. Advertisers do run recall "stars." The two high-end data points in the IRI data (see Exhibit 6) showed more sales impact than occurred for data points residing in the 80 to 160 index score range on recall. However, we must add that advertising people have long discussed, anecdotally, instances of high recall-scoring campaigns which have not been effective. This "Twilight Zone" of recall would seem to be a propitious area for future research.

The Interaction of Persuasion and Recall. Our automobile analogy, recall as the ignition system and persuasion as the fuel system, calls attention to the fact that these two elements of "how advertising works" do not work in isolation of either each other or of additional measures which may enter a multiple-measures model. Nor should we stop by inferring that recall functions solely to ignite the process. Both Haley and Baldinger (1991) and this author (Dubow, 1984) have suggested that (above threshold) recall scores may act as a multiplier of persuasion scores—but neither of the data sets are currently available for public scrutiny. This would seem to be an area ripe for further investigation.

Much research addressing the interaction of measures has already been done, and continues, in the academic literature on advertising. While we have limited our treatment of the issue to the use of sales as a dependent variable, two reviews published since "Not Recall" have covered a far wider range of measures and interrelationships (Stewart, Pechmann et al., 1985; Thorson, 1990).

Depth of Recall. The depth of recall issue is intriguing. The reason for our industry's attention to the proven-recall measure comes from the original P&G model wherein recall was designed to assess how well simple, rational measures registered with consumers. This called for obtaining verbatim records of copy playback in order to determine whether the respondent "got the message." However, as advertising messages broadened, with the inclusion of more subtlety and more emotional messages, recall hawks may have fallen too subject to the "law of the tool" and been far too slow to switch to other, better measures within the recall hierarchy.

As the types of things advertising has come to communicate have evolved, we need to ferret out which aspects are better addressed by recall, and by which recall measures, and which are better left to other measures. This, too, seems to be an area ripe for future research opportunities. It is the very essence of the multiple-measures model.

Applied Implications. There are also several applied implications worthy of our attention. These derive largely from the idea of a threshold function.

Decision-making Criteria. Using the linear assumption for recall implies that any score above zero has some impact. Under this model a manager can compare the scores for two commercials and choose the higher. This, in fact, is the model used in the analysis of the data in the CRVP—choosing among pairs. However, if the threshold model is valid, a manager laboring under the linear assumption might, in fact, be choosing among two impactless commercials. She or he would end up wasting not only the sunk costs of producing the advertising but also the (usually multiples) higher media costs which would be placed behind the advertising.

Furthermore, the reality is probably that initial losers are not discarded but left on the shelf to fight again and possibly air after being matched with a still weaker contender. Better that such commercials, if below threshold, be identified and either discarded or reworked.

15-Second versus 30-Second Commercials. If the recall threshold occurs somewhere close to an index of 80 against the norm, what is the implication for 15-second commercials? The conventional wisdom holds that :15s are 70 percent as effective as :30s—but that is a misinterpretation based on the fact that the 15-second norm is numerically 70 percent of the 30-second norm. Under the curvilinear assumption, with a threshold not much below an index of 80, a different interpretation occurs—that the average 15-second commercial is ineffective.

There seems to have been little or no research reported with regard to whether :15s are an effective form of advertising, offering another interesting arena of investigation. Until that occurs, advertisers might want to test their :15s for recall and be hesitant to use any that do not index 80 or higher against 30-second norms.

Which Norms? Category or system? If your experiences are like mine, most advertisers and agencies use category norms. The logic is that brands compete against brands and their agencies, thereby, compete against competitors' agencies. But there is an intervening variable! Before a commercial can go up against a competitor's commercial in consumers' minds, it has to go up against *all* commercials to get into consumers' minds. The battle for the share of minds is not category specific.

This difference has implications for both categories which have higher-than-system norms and ones with lower-than-system norms. For the former, which may thereby be operating in the flat portion of the recall function, there is the risk of rejecting commercials which may have lower-than-category-norm scores, but which may be just as effective. For the latter, the risk is greater; it may be that even category-normative commercials fall in the section of the curve where effectiveness drops precipitously and that barely subnormative commercials in these low-advertising-recall categories are without impact.

Admittedly, this question of "Which norm?" is a bit speculative, but given the media dollars that go behind any given commercial, were I in a low-advertising-recall category, I'd sure want to find out if I was using the right norms.

Recall Vindicated

Gibson closed his article with his title words: Not Recall. A decade later, armed with more complete information that suggests that recall *leads* (comes first, not "dominates") the advertising impact process, I shall do the same. Recall Revisited: Recall Redux! ■

JOEL S. DUBOW is an assistant professor of food marketing at Saint Joseph's University and marketing research associate with The Canaan Parish Group. Prior to that he spent 16 years fighting "The Cola Wars" in the employ of Coca-Cola USA, serving the last 13 of those years as the company's manager of communications research. He holds a Ph.D. in psychology from the University of Tennessee.

References

Achenbaum, Alvin A.; Russell I. Haley; and Ronald Gatty. "On-Air vs. In-Home Testing of TV Commercials." *Journal of Advertising Research* 7, 4 (1967): 15–19.

Advertising Research Service. "Factors Affecting Measurement of Related Recall." Evansville, IN: research systems corporation, 1979, 1980.

———. "Reliability: ARS Persuasion and Diagnostic Measures." Evansville, IN: research systems corporation, 1991.

ASI Market Research Inc. "Test-

Retest Reliability: ASI Recall." New York, NY, December 1991.

Baumwoll, Joel P. "Recall Testing Can be Dangerous to Your Brand's Health." Paper presented at the First Pan-Corporate Marketing Research Conference of General Mills, Inc. Alexandria, MN, 1978, as reported by Gibson (1983).

Burke Marketing Research, Inc. "Day-After Recall Television Commercial Testing." Cincinnati, OH, 1975, 1976, 1980.

Clancy, Kevin J., and David M. Kweskin. "TV Commercial Recall Correlates." *Journal of Advertising Research* 22, 2 (1971): 18–20.

———, and Lyman E. Ostlund. "Commercial Effectiveness Measures." *Journal of Advertising Research* 16, 1 (1976): 29–34.

Dubow, Joel S. "Starting a Copy Evaluation System from Scratch: Building a Better Wheel." In *Transcript Proceedings of the First Annual Advertising Research Foundation Copy Research Workshop.* New York: Advertising Research Foundation, 1984.

———. "The Benefits of a Standardized Systems Approach to Copy Research." In *Transcript Proceedings of the Third Annual Advertising Research Foundation Copy Research Workshop.* New York: Advertising Research Foundation, 1986.

———. "Recall First—But Not Recall Alone." In *Transcript Proceedings of the Ninth Annual Advertising Research Foundation Copy Research Workshop.* New York: Advertising Research Foundation, 1992.

———, and Roy G. Stout. "Redefining Communication." Paper presented to the Association of National Advertisers Advertising Research Workshop. New York, December 1983.

Gibson, Lawrence D. "Not Recall." *Journal of Advertising Research* 23, 1 (1983): 39–46.

Greene, William F. "Copy Research Validation." Paper presented to the Advertising Research Foundation Copy Research Validation Key Issues Workshop. New York, November 1981.

Haley, Russell I., and Allan L. Baldinger. "The ARF Copy Research Validity Project." *Journal of Advertising Research* 31, 2 (1991): 11–32.

Heller, Harry E. "The Ostrich and the Copy Researcher: A Comparative Analysis." A talk delivered at the American Marketing Association New York Chapter luncheon meeting of the Advertising Effectiveness Research Group. New York, December 1971.

Information Resources Inc. "How Advertising Works." Chicago: Information Resources, Inc., 1991.

Lamar, Charles M. "How We Know Our Copy Testing Is Valid." Paper presented to the Association for Consumer Research Twelfth Annual Conference. St. Louis, October 1981.

Ross, Harold L. "Recall versus Persuasion: An Answer." *Journal of Advertising Research* 22, 1 (1982): 13–16.

Schulman, Art. "On-Air Recall by Time of Day." *Journal of Advertising Research* 12, 1 (1972): 21–23.

Silk, Alvin J. "Test-Retest Correlations and the Reliability of Copy Testing." *Journal of Marketing Research* 14, 4 (1977): 476–86.

Stewart, David W.; David H. Furse; and Randall P. Kozak. "A Guide to Commercial Copy Testing Services." *Current Issues & Research in Advertising* 6, 1 (1983): 1–44.

———; Connie Pechmann; Srinivasan Ratneshwar; Jon Stroud; and Beverly Bryant. "Methodological and Theoretical Foundations of Advertising Copytesting: A Review." *Current Issues and Research in Advertising* 8, 2 (1985): 1–74.

Stout, Roy G. "Copy Testing Is Only Part of Advertising Research." Paper presented to the Association for Consumer Research Twelfth Annual Conference. St. Louis, October 1981.

TeleResearch, Inc. "More about the Use and Limitations of Recall Scores." *Tele/Scope* 3, 4 (1970): 1–8.

Thorson, Esther. "Consumer Processing of Advertising." *Current Issues and Research in Advertising* 12, 2 (1990): 197–230.

Walker, David, and Michael F. von Gonten. "Explaining Related Recall Outcomes: New Answers from a Better Model." *Journal of Advertising Research* 29, 3 (1989): 11–21.

Young, Shirley. "Copy Testing without Magic Numbers." *Journal of Advertising Research* 12, 1 (1972): 3–12.

Yuspeh, Sonia. "The Medium Versus the Message: The Effects of Program Environment on the Performance of Commercials." Paper presented to the Tenth Annual Attitude Research Conference of the American Marketing Association. Hilton Head, SC, February 1979.