An Analysis of Real World TV Advertising Tests: A 15-Year Update

YE HU
University of Houston
yehu@uh.edu

LEONARD M. LODISH
The Wharton School of the University of Pennsylvania
lodish@wharton.upenn.edu

ABBA M. KRIEGER
The Wharton School of the University of Pennsylvania
krieger@wharton.upenn.edu

An analysis is performed on the results of 241 real world TV advertising tests conducted by Information Resources, Inc. between 1989 and 2003 to partially update the findings of Lodish et al. [*Journal of Marketing Research* 32, 2 (1995): 125–39]. Two types of market test results, BehaviorScan and Matched-Market, are analyzed. Overall, the improvement of TV advertising sales effectiveness because of media weight increase is significantly larger than zero for established products, which is different from Lodish et al.'s finding. A further analysis indicates that such significance is mainly driven by more recent tests. A comparison between the new results and Lodish et al. reveals a significant increase in the average advertising effectiveness for tests completed after 1995. The new data still suggest (as did the original data) that it is of great managerial interest to identify advertising effectiveness before launching advertising campaigns.

IT HAS BEEN MORE THAN 10 years since Lodish et al. (1995a) published their meta-analytical study of IRI (Information Resources Inc.) BEHAVIORSCAN®'s (BSCAN) split cable in-market TV advertising experiments. This article partially updates that analysis and provides further managerial implications of the new results. [This study is a "partial" update of Lodish et al. (1995a) because of limited resources. In particular, the study by Lodish et al. (1995a) included detailed survey information about the advertising, marketing, competitive strategy, and tactics for each of the advertising campaigns, and extensive product competitive situations during each of the BSCAN tests. Such information cost over $1 million to gather. These resources were not available for this analysis.] From the original analysis, Lodish and colleagues reached a number of conclusions: some reinforced common beliefs about TV advertising—such as new-product TV advertising has greater impact on sales than TV advertising for established products; others were surprising and counter to conventional wisdom in the advertising industry—that increasing advertising budgets does not increase sales in general, and that there is not a monotonic relationship between increasing TV advertising and increasing sales. The in-market experiments used in Lodish et al. (1995a)'s study were completed between 1982 and 1988. Changes (e.g., audience segmentation; a typical TV advertisement used to last 1 minute, now it is usually 30 seconds in length; increased media fragmentation) during the past decade and a half might affect both the effectiveness and the managerial implication of weight increases of TV advertising.

During the 1990s, IRI introduced an additional in-market test method called Matched-Market (MMT) test, which allows for differential test and control conditions in two geographically distinct markets that are matched on many factors including weekly sales and market share of the test brands. A more detailed comparison between procedures of BSCAN and MMT is provided in the Appendix. In this study, we update the meta-analysis study by Lodish and colleagues on BSCAN and also report and analyze results from the MMT tests. Our analysis is based on a new set of 241 TV

DOI: 10.2501/S0021849907070353

advertising tests completed by IRI between 1989 and 2003.

The article is organized as follows. We first briefly review the relevant literature on meta-analyses of advertising and TV advertising weight change, with an emphasis on new publications post-1995. We then perform the analysis and compare the results from the new tests with the results from Lodish et al. (1995a). Based on the elasticity results, we further analyze short-term as well as long-term return of advertising investment for typical consumer package goods. We conclude with the implications of the old and new set of tests for marketing/advertising managers and researchers.

LITERATURE REVIEW

Recent meta-analysis of advertising has generated findings that bear on various advertising characteristics. For instance, Abernethy and Franke (1996) find, contrary to claims by critics, that advertising does provide information to the consumers. The amount of information varies across media. Grewal et al. (1997)'s analysis shows that comparative advertisements are more effective than noncomparative advertisements in generating attention, message, and brand awareness; levels of message processing; favorable sponsored brand attitudes; and increased purchase intentions and purchase behaviors. They also discover that comparative advertisements evoke lower source believability and a less favorable attitude toward the advertisement. Brown, Homer, and Inman (1998) find that positive and negative feelings have contingently asymmetrical effects on advertising responses.

The literature also addresses the differential effect of certain advertising strategies across different markets. Chandy, Tellis, MacInnis, and Thaivanich (2001) study how advertising cues affect consumer behavior in new versus well-established markets. Based on data of a toll free referral service, their analysis indicates that argument-based appeals, expert sources, and negatively framed messages are particularly effective in new markets. They also find that emotion-based appeals and positively framed messages are more effective in older markets than in newer markets. Erdem and Sun (2002) investigate and find evidence for advertising and sales promotion spillover effects for umbrella brands in frequently purchased packaged product categories. Their study provides empirical evidence for the uncertainty-reducing role of advertising across categories for umbrella brands.

A thorough review on advertising by Vakratsas and Ambler (1999) shows no additional research on the short-term advertising elasticity published, at least by the time the article was written. The recent literature that is more closely related to our research on the short-term sales impact of TV advertising *media weight* takes a "micro" perspective, focusing on the specifics of the advertisements themselves. For instance, MacInnis, Rao, and Weiss (2002) assess the conditions under which increased media weight of advertising helps sales. They demonstrate, through analyses of a series of existing experiments, that affect-based executional cues in advertising tend to increase media-weight-induced sales in established product categories. They also find that it helps to combine the evocation of positive (and no negative) feelings in order to increase the effectiveness of media weight. Naik, Mantrala, and Sawyer (1998) have developed and empirically tested a media planning model based on the dynamics of advertising wear-out. Tellis, Chandy, and Thaivanich (2000) model the short-term effect of TV advertising for a toll-free service to empirically answer the questions of which advertisement works and, when, where, and how often it works. Steenkamp, Nijs, Hanssens, and Dekimpe (2005) find that the ultimate impact of most promotion and advertising campaigns depends primarily on the nature of consumer response, not the vigilance of competitors. On the other hand, research on the long-term effect of advertising mainly investigates how much (or whether) it indeed decreases consumers' price sensitivity (Dekimpe and Hanssens, 1999; Jedidi, Mela, and Gupta, 1999; Kaul and Wittink, 1995; Mela, Gupta, and Lehmann, 1997). Due to the rare availability and difficulty in acquiring expensive clean real world advertising weight experimental test data, the recent research above depends either on controlled lab experiments or more sophisticated econometric methodologies such as time series analysis. Our new data from IRI, similar to Lodish et al. (1995a), are controlled "in-market experiments" that provide us with a relatively clean set of before-and-after results for the analysis.

RESULTS AND IMPLICATIONS

Each TV advertising test conducted by IRI is composed of a test group and a control group. Before the test starts, these two groups receive the same base level of advertising treatment for a period of usually 52 weeks. During the test period, the test and control groups are given different treatment levels of TV advertising exposures. At the end of the test period, the percentage sales difference (in volume) between the test and the control groups, after adjusting for covariates that account for group differences, is reported as the "Adjusted Volume Effect" (AVE). [The raw test results are the volume sales of the test brand between the test and the control groups. Because certain marketing mix variables in the test market, such as feature, display, and price changes, cannot be experimentally controlled, the raw test

results have to be adjusted for these covariates using ANCOVA in order to isolate the actual advertising influence. AVE represents the percentage sales difference (in volume) between the test and the control groups after such adjustment. More information on the covariates used in the ANCOVA is provided in the Appendix.] AVE measures the sales effectiveness of advertisements due to media weight change. For each test, the AVE as well as its *p*-value are reported. These two inputs are used to recover the standard deviation of AVE *within* each test (standard deviation = AVE/$\Phi^{-1}(1-p)$, where $\Phi(.)$ is the cumulative density function of a standard normal distribution), which reflects the variability of AVE within each test.

Two kinds of advertising weight changes are applied in both BSCAN and MMT: The test group receives base-level advertising exposure and the control group receives no advertising exposure—this is called an "Ad/No Ad Test" (Ad/NoAd); or both the test and the control groups have advertising exposures (at different media weight levels)—this is called a "Weight Test" (WEIGHT). Additionally, we observe fairly dispersed AVE values across different types of tests (BSCAN versus MMT, Ad/NoAd versus WEIGHT) as well as over time. We therefore split the tests further into two time periods: those completed on or before December 31, 1995 versus those completed after December 31, 1995 (denoted as BEFORE95 and AFTER95, respectively. The number of BEFORE95 BSCAN tests is roughly equal to that of AFTER95). In this section, we report analyses performed on three sets of dependent variables: AVE, significance level and variability of AVE, and advertising elasticity, where advertising elasticity is defined as the percentage change of volume sales divided by the percentage change of advertising weight, i.e., the AVE divided by percentage change of advertising weight:

$$\text{Advertising Elasticity} = \frac{\Delta\text{Volume}/\text{Volume}}{\Delta\text{AdWeight}/\text{AdWeight}}.$$

Adjusted volume effect

First, we test whether the AVE's are significantly different from zero. Due to the small sample size of new-product tests, our further analyses are based on 241 established products. [There are only 12 tests on new products (AVE: Mean = 0.256, Standard Error of Average AVE across tests = 0.072, $p < 0.05$, which reconfirms the finding by Lodish et al. (1995a) that new-product advertising has significant positive effect on sales).] We use a *t*-test, where the *t*-statistic is calculated using column 3 (Average AVE) and column 5 (Standard Error of Average AVE across tests) in Table 1 (note that 31 tests with missing significance levels are excluded). The average AVE of both BSCAN and MMT tests is significantly larger than zero (*p*-value < 0.001). [We are faced with the following situation. The AVE for each test A_i is measured with error. The error (because the ANCOVA coefficient is unbiased) has mean zero and variance that is reported for each test. These variances, τ_i^2 (or the square root of the same), is what we call standard deviation *within* test. In doing the *t*-test we assume that the AVEs across tests are random draws from a distribution with a common variance σ^2. Hence $\text{Var}(A_i) = \tau_i^2 + \sigma^2$. We need to compute the standard error of the average. The variance for the average is then

$$Var\left[\frac{A_1 + \cdots + A_n}{n}\right] = \frac{\left[n\sigma^2 + \sum_i \tau_i^2\right]}{n^2},$$

where n is the number of tests. Note that the standard deviation across tests implicitly takes into account both with test and across tests variability (column 4 in Table 1).] It shows that for established products, advertising weight increase (including both WEIGHT and Ad/NoAd) leads on average to statistically significant sales increases. We use the BSCAN WEIGHT test results from our new data to make comparisons with the data (N = 206) used by Lodish et al. (1995a) because all tests in their analysis are of this type. Overall, the average AVE of the new data is significantly higher than that of Lodish et al. (1995a) ($t = 2.38$, $p = 0.018$). A further analysis shows that the observed difference is mainly driven by the test results after 1995: the new tests before 1995 are not significantly different from the corresponding tests in Lodish et al. (1995a) ($t = 0.76$, $p = 0.225$); while the

TABLE 1
Overall Adjusted Volume Effect of Established Products

Tests (1)	*N* (2)	Average of Adjusted Volume Effect (Average AVE) (3)	(Average of) Standard Deviation of AVE *Within* Each Test (4)	Standard Error of Average AVE (Across Tests) (5)
All tests	210	0.104	0.102	0.011
BSCAN	89	0.140	0.172	0.024
MMT	121	0.078	0.050	0.007

For established products, advertising weight increase leads on average to statistically significant sales increases.

tests after 1995 have significantly higher average AVE than those in Lodish et al. (1995a) ($t = 2.68$, $p < 0.01$). This shows that on average advertising weight change has become more effective in recent years.

The relationship between GRP% change and AVE of BSCAN and MMT WEIGHT tests is shown as scattered diagrams in Figure 1. Overall, BSCAN WEIGHT tests completed after 1995 (triangles in Figure 1) tend to have higher AVE with the same GRP% change.

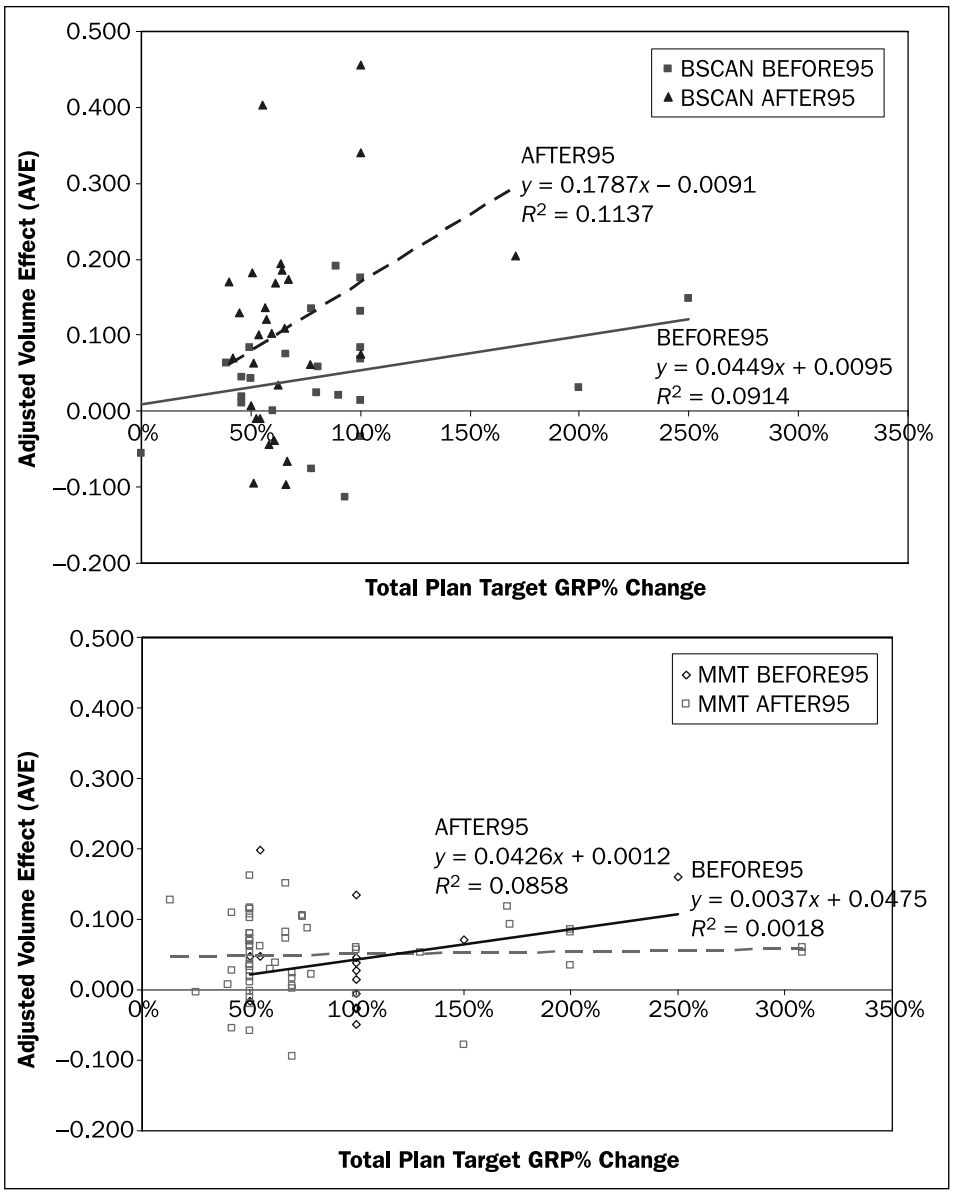

Figure 1 Percent Volume Increase Associated with Percent Target GRP Increase for Established Product Weight Tests

To investigate how AVE varies across different test conditions, we use ANOVA where the dependent variable is AVE and the three factors are test type (BSCAN versus MMT), test weight type (Ad/NoAd versus WEIGHT), and time of the test (BEFORE95 versus AFTER95), as well as all three two-way interactions. The overall effect is significant ($F = 6.41$, $p < 0.001$). The interactions as well as the main effects are significant (Table 2). Specifically, there is a tendency for AVE to increase for the following groups: BSCAN tests that are Ad/NoAd; BSCAN tests that were completed after 1995; Ad/NoAd tests that were completed after 1995.

To summarize results reported in Table 2, first, Ad/NoAd tests have an estimated increase in AVE of 0.061 compared to weight tests. This result is consistent with the findings in previous research (e.g., Lodish et al., 1995a) that there exists diminishing marginal effectiveness of advertisement expense. This can be seen because the change of advertising exposure from zero to nonzero (Ad/NoAd case) on average generates higher AVE increase than in the weight test cases, i.e., going from no advertising to some advertising has greater impact than going from some advertising to more advertising. Second, advertisements that are executed after 1995 have an estimated increase in AVE of 0.079 as compared to earlier tests. This supports the general conclusion of this study that advertisement effectiveness is tending to increase in recent times. Finally, BSCAN tests have an estimated increase in AVE of 0.068 as compared to MMT tests. This is probably due to the difference between the two test procedures, as described in the Appendix. The ANOVA above is based on the assumption that the residuals follow a normal distribution. However, a fair level of skewness is

TABLE 2
ANOVA Test of Adjusted Volume Effect

	Effect			
Variable	**Mean**	**Standard Error**	***t*-Ratio**	***p*-Value**
BSCAN[a]	0.0342	0.0111	3.08	0.0023
Ad/NoAd[b]	0.0307	0.0109	2.82	0.0053
BEFORE95[c]	−0.0394	0.0111	−3.55	0.0005
BSCAN ∗ (Ad/NoAd)	0.0279	0.0110	2.53	0.0119
BSCAN ∗ BEFORE95	−0.0279	0.0107	−2.60	0.0100
(Ad/NoAd) ∗ BEFORE95	−0.0240	0.0111	−2.17	0.0313

[a]*BSCAN = 1 if a test is a BSCAN test; −1 otherwise.*
[b]*Ad/NoAd = 1 if a test is an Ad/NoAd Test; −1 if a test is a Weight Test.*
[c]*BEFORE95 = 1 if a test was complete before 1995; −1 otherwise.*

observed in the normal quantile plot of residuals (Figure 2). For this reason, we also use a nonparametric approach (Wilcoxon/Kruskal-Wallis Test) and report the results in Table 3. The overall effect is again significant ($\chi^2 = 14.81$, $df = 7$, $p = 0.039$). The tests that contribute most by having very high AVE are those that are Ad/NoAd, BSCAN, and are after 1995 (denoted by BSCAN_Ad/NoAd_AFTER95 in Table 3). This is consistent with the ANOVA results.

Significance level and variability of AVE

We observe *p*-values for 210 tests. If the effectiveness of these TV advertising tests is random and these tests in general do not lead to sales increase, these *p*-values would have a uniform distribution. If advertising effect exists in these tests, the *p*-values should be stochastically smaller (i.e., more significant) than a uniform distribution on the unit interval. A Kolmogorov-Smirnov (KS) test (cf., Pearson and Hartley, 1976, pp. 117–119 and Table 54) is performed to see whether there is a greater-than-chance likelihood of achieving lower *p*-values (thus a larger number of tests with significant AVE). [This is a more appropriate method than Dutka (1984) adopted by Lodish et al. (1995a).] KS performed for all 210 cases yields a value of 6.59. Similarly, the data divided into the tests prior to 1995 ($N = 83$) and post-1995 ($N = 127$) have KS values of 3.92 and 5.45, respectively. All three values exceed 1.518 (the cutoff value when $p = 0.01$); hence we can reject the assumption that *p*-values are uniform in favor of a tendency to have smaller (i.e., more significant) *p*-values. An alternative approach is to consider the tests in a multiple comparison framework and test the aggregate null hypothesis that none of the tests are significant. A modern approach for dealing with this issue (preferred to the perhaps more recognized Bonferroni correction) is the linear step-up procedure (cf., Hochberg and Tamhane, 1987). Specifically, the n (number of tests) *p*-values are ordered from the smallest to the largest. The ith smallest *p*-value is tested at $0.05i/n$. Let k be the largest number such that the kth smallest *p*-value does not exceed $0.05k/n$. The k tests corresponding to the k smallest *p*-values are rejected. The number of tests that would be rejected out of the 210 tests are 63, thereby demonstrating that there are tests that are significant, which is consistent with the results of the KS test. Of the 63 rejected tests, about the same fraction (24 out of 83 before 1995; 39 out of 127 after 1995) or about 30 percent come from the two time periods. From the above two tests, we conclude that

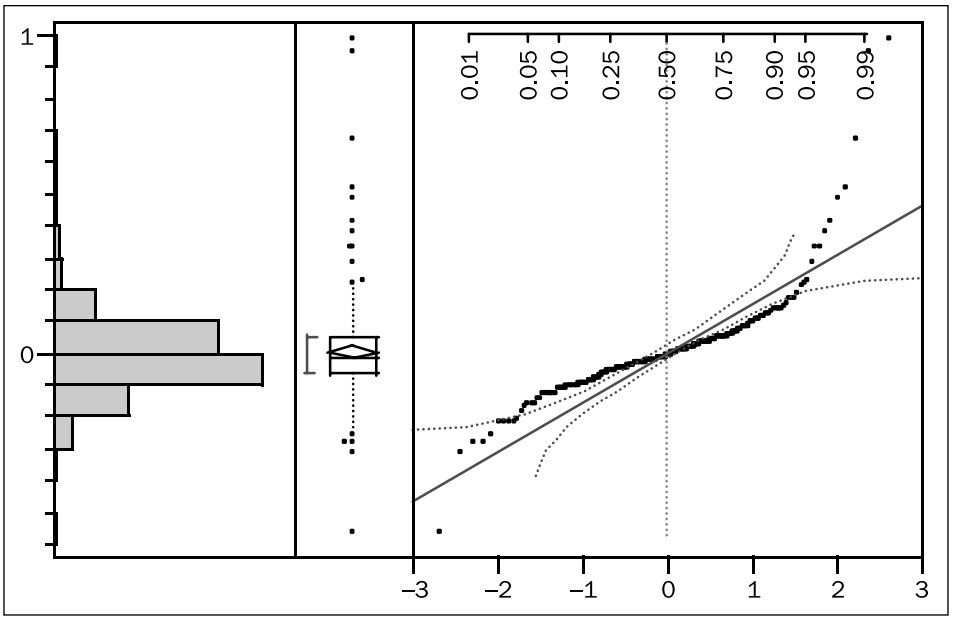

Figure 2 Normal Quantile Plot of AVE ANOVA Residuals

TABLE 3
A Nonparametric Comparison of Adjusted Volume Effect (Wilcoxon/Kruskal-Wallis Tests—Rank Sums)

Factor	*N*	Score Sum	Score Mean	(Mean-Mean0)/Std0
BSCAN_Ad/NoAd_BEFORE95	21	2333	111.1	−0.680
BSCAN_Ad/NoAd_AFTER95	17	2806	165.1	2.701
BSCAN_WEIGHT_BEFORE95	29	3170	109.3	−0.963
BSCAN_WEIGHT_AFTER95	31	4344	140.1	1.634
MMT_Ad/NoAd_BEFORE95	16	2131	127.0	0.353
MMT_Ad/NoAd_AFTER95	29	3870	133.4	1.022
MMT_WEIGHT_BEFORE95	25	2600	104.0	−1.286
MMT_WEIGHT_AFTER95	73	8008	109.7	−1.658

there is significantly above random chance that the test results are significant.

We also compare the variability in the AVE across different time periods. This variability is an indicator of how uncertain the sales effectiveness of a typical advertising test campaign is. The classical F-test is not relevant for these data, as the observations are highly skewed; it is well known that the F-test is very sensitive to the assumption of normality. For this reason, we use an Ansari-Bradley nonparametric dispersion test (cf., Hollander and Wolfe, 1999). We restrict attention to BSCAN weight tests as these are the only tests that were available from the original Lodish et al. (1995a) paper. The null hypothesis is that there is no difference in dispersion between the Lodish et al. data set and the new tests or its two subsets (before and after 1995). All three comparisons produce *p*-values that exceed 0.05. Furthermore, the dispersion of the tests before 1995 is less than the dispersion of the tests post-1995 ($p < 0.001$). Generally speaking, the variation in the test results has not decreased (or even increased) post-1995 comparing to before 1995.

Elasticity

Advertising elasticity is another measure of the effectiveness of TV advertising weight change; its advantage, comparing to AVE, is that the amount of weight change is also taken into consideration. In this section, we repeat the earlier analysis of AVE, but use elasticity instead as the dependent variable. Our analysis of advertising elasticity is only based on the weight tests ($N = 127$). Obviously, the elasticity of Ad/NoAd tests is not well defined given that the base level of advertising expense is 0. We use a *t*-test to test whether the average elasticity is larger than zero, where the *t*-statistic is calculated using column 3 (Average Elasticity) and column 5 (Standard Error of Average Elasticity across tests) in Table 4. The average elasticity of both BSCAN and MMT tests are significantly larger than zero (*p*-value < 0.05), which is consistent with the earlier analysis of AVE and different from Lodish et al. (1995a). Furthermore, we compare the elasticity of our new data with that of Lodish et al. (1995a). [Out of the 206 BSCAN WEIGHT tests in Lodish et al. (1995a), 44 tests have advertising elasticity available. A *t*-test of the average AVE between the 206 tests and the 44 with elasticity available shows they are not significantly different (average AVE of the two groups = 0.0397 and 0.0358, respectively, *p*-value = 0.88).] A Wilcoxon/Kruskal-Wallis nonparametric test is used because of skewness in the data. The overall comparison is marginally significant ($\chi^2 = 5.62$, $df = 2$, $p = 0.06$) and the elasticity of new tests completed after 1995 turns out to be significant (Table 5). The conclusion that BSCAN tests post-1995 tend to have higher elasticity is consistent with the analysis of AVE.

To analyze the effect of test factors on elasticity, an ANOVA is performed where the dependent variable is the elasticity and the two factors are test type (BSCAN versus MMT) and time of the test (BEFORE95 versus AFTER95), as well as the interaction between test type and time. This test is insignificant ($F = 1.166$, *p*-value = 0.326). Because the residuals from ANOVA are again skewed, we performed a Wilcoxon/Kruskal-Wallis nonparametric test. This test is marginally significant ($\chi^2 = 7.41$, $df = 3$, $p = 0.06$).

Generally speaking, the variation in the test results has not decreased (or even increased) post-1995 comparing to before 1995, . . . an indication that the heterogeneity in the advertising response function should not be ignored.

TABLE 4
Overall Elasticity of Weight Tests

Tests (1)	*N* (2)	Average Elasticity (3)	(Average of) Standard Deviation of Elasticity *Within* Each Test (4)	Standard Error of Average Elasticity (Across Tests) (5)
All Tests	127	0.113	0.139	0.021
BSCAN	52	0.120	0.214	0.023
MMT	75	0.108	0.085	0.031

TABLE 5
A Nonparametric Comparison of Elasticity between New Data and Lodish et al. (1995a), WEIGHT Tests Only (Wilcoxon/Kruskal-Wallis Tests—Rank Sums)

Factor	*N*	Score Sum	Score Mean	(Mean-Mean0)/Std0
Lodish et al. (1995a)	44	1883	42.8	–1.842
BSCAN_WEIGHT_BEFORE95	23	1078	46.9	–0.318
BSCAN_WEIGHT_AFTER95	29	1695	58.4	2.298

The details are reported in Table 6. It appears that the groups with significantly higher elasticity are BSCAN WEIGHT tests completed after 1995.

Additionally, we use elasticity results to test whether the marginal return on advertising expense is monotonic, increasing, or decreasing. The elasticity is essentially the slope in the relationship between ln(AVE) and ln(Ad Weight Change %). A positive slope (elasticity) says that as weight increases so does the effect of the advertising increase. We related the elasticity to the size of the weight to see if the elasticity is constant. The slope is (insignificantly) negative due to the one large elasticity (Figure 3A, Coefficient of Increase Weight %: mean = −0.027, *t*-ratio = −1.46, *p*-value = 0.146). We redid the above analysis removing the one potential outlier (Figure 3B). Interestingly, because of reduction in variability the *p*-value goes down, but remains greater than 0.05. The slope remains negative and becomes marginally significant (Coefficient of Increase Weight %: mean = 0.022, *t*-ratio = −1.79, *p*-value = 0.076). Such a downward sloping relationship, consistent with our earlier analysis of AVE, indicates that there exists diminishing marginal effectiveness of advertising expense over all. Locally, however, such relationship could be more complicated because of a possible nonlinear relationship (e.g., an "S" shape) between the two.

TABLE 6
A Nonparametric Comparison of Elasticity, WEIGHT Test Only (Wilcoxon/Kruskal-Wallis Tests—Rank Sums)

Factor	*N*	Score Sum	Score Mean	(Mean-Mean0)/Std0
BSCAN_WEIGHT_BEFORE95	23	1373	59.70	–0.167
BSCAN_WEIGHT_AFTER95	29	2270	78.28	2.375
MMT_WEIGHT_BEFORE95	15	728	48.53	–1.729
MMT_WEIGHT_AFTER95	60	3757	62.62	–0.398

DISCUSSION AND CONCLUSION

In this research, we update Lodish et al. (1995a) by analyzing a new set of IRI TV advertising tests. Some interesting findings about the effectiveness of TV advertising weight changes emerge from our analysis. Specifically, recent TV advertising tests show that, unlike what Lodish et al. (1995a) have found based on earlier data, the average advertising elasticity becomes statistically significantly larger than zero. The current data do not enable us to distinguish among the potential reasons. Because company-level analysis is not realistic (the tests were conducted for 45 companies), we split the advertised products into four categories (Cleaning, N = 33; Drink, N = 6; Food, N = 144; and Health and Beauty, N = 58) to investigate whether these categories may drive the change of advertising effectiveness before and after 1995. Neither the analysis of AVE nor that of elasticity returns significant *p*-values for the product category effect. [The ANOVA of AVE (similar to the analysis report in Table 2, with category factors added) gives category

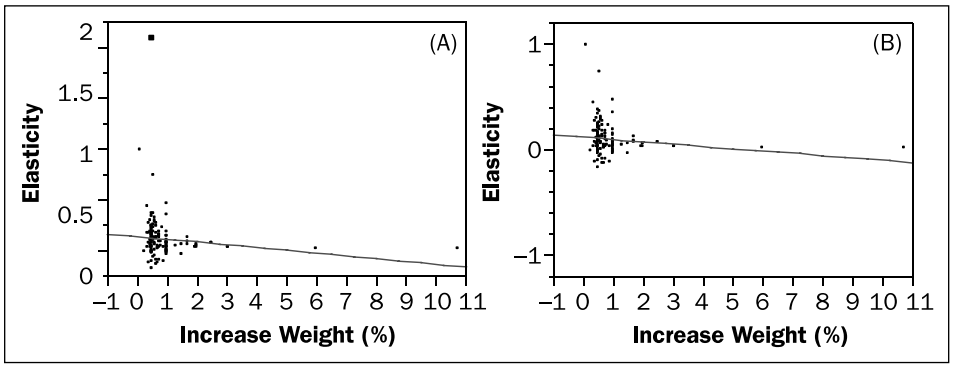

Figure 3 Diminishing Return of Advertising Weight Increase

factors a *p*-value of 0.976. The nonparametric comparison of elasticity (similar to the analysis reported in Table 6, with category factors added) gives a χ^2 of 24.69 (df = 22, p-value = 0.31).] Thus, we exclude it as a possible explanation. Of course, further investigation in this regard would be of great practical value to the advertising industry. We believe it is an interesting topic worthwhile for future research.

What are the implications of our analysis for the profitability of the TV advertising investment for a *typical* packaged good TV advertiser? [We thank Andrew Shore, a senior analyst for Global Consumer goods at Deutsche Bank Securities, for providing estimates for a typical package good. A conservative estimate of the incremental marginal contribution of revenue of a typical firm: 65 percent and of the advertising budget as percent of sales: 5 percent (including estimates for agency fees and production costs for the TV advertisements). We take a "typical" brand with $100 million in revenue and a $5 million in TV advertising expenditure.] A study of BSCAN tests by Lodish et al. (1995b) experimentally measured the long-term (2–3 years) impact of advertising weight changes. The study indicates that the short-term (1 year) sales effectiveness of advertisements with smaller AVE *p*-values ($p \leq 0.2$ in Lodish et al., 1995b) is approximately doubled over the next 2 years because of the long-term impacts of the test advertising treatment. [We realize that $p = 0.2$ is not a conventional criterion for statistical significance. In this demonstration, we split the new tests into separate groups using $p = 0.2$ as Lodish et al. (1995b) did in order to apply their findings. Note that all analyses in this article use $p \leq 0.05$ for statistical significance.] Those advertising tests that were less significant ($p > 0.2$) had no long-term effect; that is, if TV advertising works in the short term, its impacts are doubled over the next 2 years. If the TV advertising does not work the first year, it will not have any long-term impact. We applied these long-term

TABLE 7
"Typical" Long-term Revenue and Profit Impact of TV Advertising (BSCAN Weight Tests) for a $1 Million Advertising Increase

				Short-Term (1-year) Impact		Long-Term (Including 2- and 3-year) Impact	
BSCAN Weight Tests	Category	*N*	Average Elasticity	Increased Revenue per TV Advertising $Million Increase ($Million)	Profit Change ($Million)	Revenue Change ($Million)	Profit Change ($Million)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Before 1995	$p \leq 0.2$	9	0.136	2.72	0.77	5.44	2.54
	$p > 0.2$	14	0.020	0.40	–0.74	0.40	–0.74
	Weighted sum	23	0.065	1.31	–0.15	2.37	0.54
After 1995	$p \leq 0.2$	13	0.303	6.06	2.94	12.12	6.88
	$p > 0.2$	16	0.050	1.00	–0.35	1.00	–0.35
	Weighted sum	29	0.163	3.27	1.12	5.98	2.89

findings to the new BSCAN test results (Table 7). The long-term estimates in Table 7 assume that the long-term analysis conducted by Lodish et al. (1995b) is still valid for BSCAN experiments even though the tests analyzed were from the 1980s. There is no other, newer experimental estimate to apply. The reader can interpret the long-term estimates with an appropriate "grain of salt." For concreteness, we illustrate the impact of an additional $1 million in TV advertising. Column 5 is estimated by taking the average of elasticity in column 4 and applying it to the 20 percent increase in TV advertising to estimate the short-term revenue increase for an extra million dollars of TV advertising. Column 6 then multiplies this incremental revenue by 0.65 and subtracts the $1 million cost of the incremental TV advertising to obtain short-term profit change. Column 7 uses only the tests with $p \leq 0.2$ and doubles their average short-term sales increase to reflect the long-term results cited above. These estimates put into perspective the increase in sales effectiveness we see after 1995. The tests with $p \leq 0.2$ have a short-term average profit return on advertising investment of over 200 percent just during their first year. If the Lodish et al. (1995b) results still hold, the return on profit impact over 3 years is almost 700 percent. On the other hand, even in the after-1995 period, those test campaigns with $p > 0.2$ had average negative profit returns on advertising investment of −35 percent. Overall, the "weighted sums" of the long-term returns of all of these tests are 54 percent before 1995 and 289 percent after 1995. For a test with $p > 0.2$, the break-even elasticity is 0.077; for a test with $p \leq 0.2$, the break-even elasticity is much lower, 0.026. Obviously, being able to identify advertising campaigns with more significant effect (small p-values) gives an edge in returns on advertising expense. The projected short- and long-term profitability of tests in our dataset is illustrated in Figure 4.

Additionally, we find a significant difference between the group of Weight tests and Ad/NoAd tests and between tests that were conducted before and those that were conducted after 1995. The results are of importance to the practices of the advertising industry. They

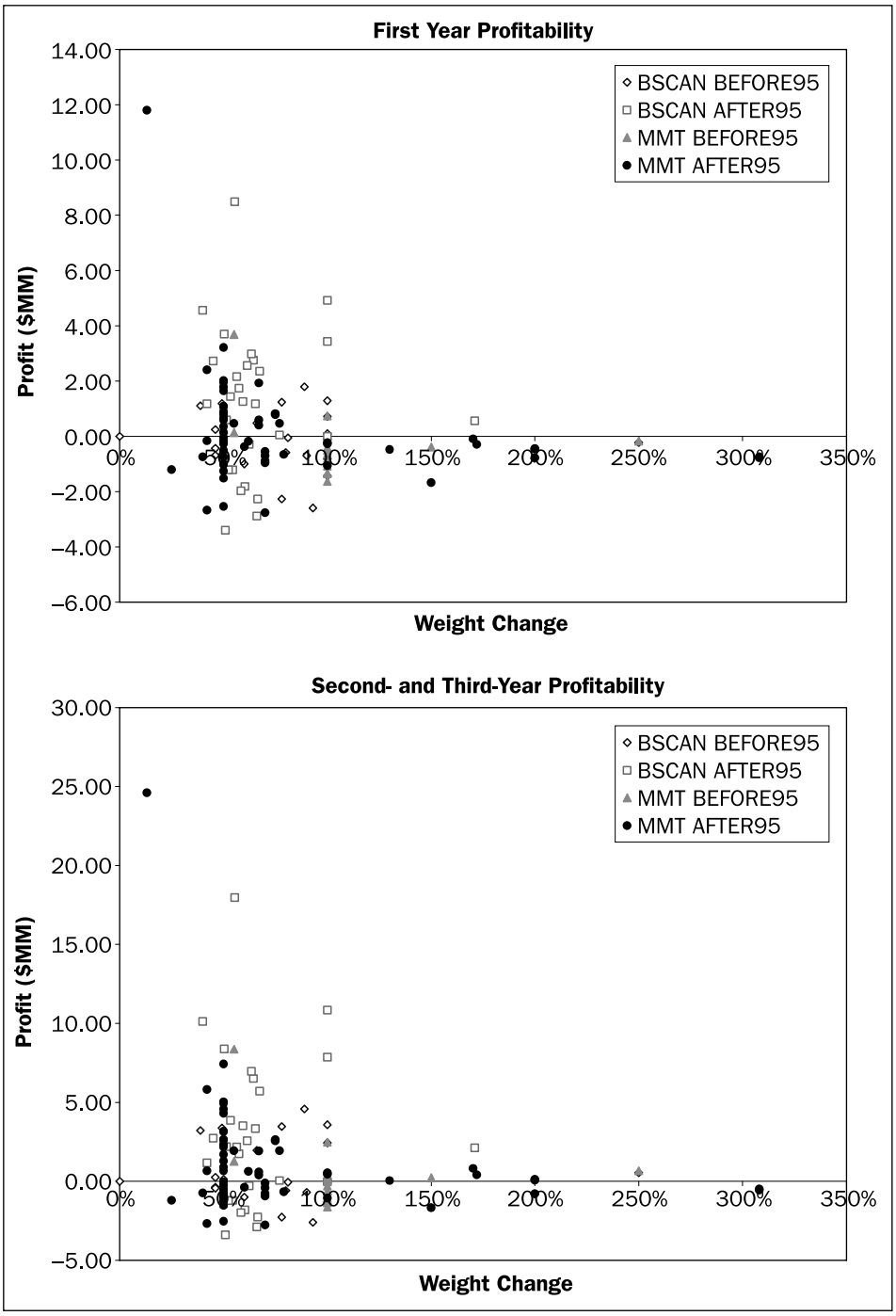

Figure 4 Projection of Advertising Campaign Profitability (First Year and Long Term)

We find a significant difference between . . . tests that were conducted before and those that were conducted after 1995.

demonstrate that regular advertising weight change differs from Ad/NoAd situations. The increase of the effectiveness of advertising weight change over time is encouraging. Given the diminishing return to increase in advertising expense suggested by the results, instead of blindly increasing advertising weight, it is important to take the return of investment into consideration to improve the return on the expense of advertising.

Moreover, we find that there is large variability in advertising effectiveness within each test (measured by the AVE standard deviation within each test), an indication that the heterogeneity in the advertising response function should not be ignored in academic research. In accordance with our conclusion, Chintagunta and Vilcassim (1995) find that the advertising response coefficients of two competing beer brands are distinctively different. A recent study by Vakratsas, Feinberg, Bass, and Kalyanaram (2004) suggests that dynamic, environmental, competitive, and brand-specific factors can influence advertising effectiveness and, subsequently, estimated results of the advertising response function. The variability translates into a large uncertainty about the sales effectiveness of a TV advertising campaign when it is run. Such high variation in advertising effectiveness together with the lucrative long-term return of advertisements with low *p*-values as we have discussed earlier indicate that it is very important to balance the choice between running TV advertisements at a fast pace versus generating, testing, and selecting the "right" TV advertising copy. A Bayesian decision maker would not support the typical way most companies are managing this uncertainty in advertising sales impact of different campaigns.

We demonstrate, through both short- and long-term analysis of a "typical" package good, that it is very important to balance the choice between running TV advertisements at a fast pace versus generating, testing, and selecting the "right" TV advertising copy. In reality, the TV advertising world seems to be heading in the former direction. According to a recent article in *Fortune* (Leonard, 2004, p. 96), ". . . media buyers are providing the bulk of their parent companies growth, while the creatives' flounder." Moreover, ". . . Media buyers offered something that every corporate procurement executive could understand: cost savings." This phenomenon is another example of how precise data (media costs) become much more important than data (creative outcomes) that take months to measure and are not very precise. It will take courageous senior marketing managers and company CEOs to challenge the current agency structure and insist on generating (and paying well for) creative copy that works. We hope this article will shift attention to creativeness and proactively managing TV advertising sales impact variability from cost-cutting measures promoted by the media buyers and corporate purchasing agents. We also hope it will spur industry investment in development of reasonably valid, relatively quick, copy evaluation methodologies that firms will use to increase their profitability. JAR

YE HU is an assistant professor of marketing at the C.T. Bauer College of Business, University of Houston. He received a Ph.D. in marketing from The Wharton School of the University of Pennsylvania. His recent research focuses on applying discrete choice models, Bayesian statistical models, and survival models to better understand individual choice and decision making in various retailing settings such as internet auctions and gift registries. His publications have appeared in the *Journal of Marketing Research*.

LEONARD M. LODISH is the Samuel R. Harrell Professor of Marketing at The Wharton School of the University of Pennsylvania. He received a Ph.D. in marketing and operations research from Massachusetts Institute of Technology. He is also vice dean for Wharton West. Prof. Lodish has published numerous papers and books in the areas of entrepreneurial marketing, strategic and tactical marketing resource planning, marketing decision support systems, and applications in firm/marketing strategy, sales force, advertising, and promotion planning.

ABBA M. KRIEGER is the Robert Steinberg Professor of Statistics and Operations Research, Operations and Information Management, and Marketing, and a fellow of the American Statistical Association. He received a Ph.D. in statistics from Harvard University. Prof. Krieger has published numerous papers on bootstrap, grouped data, density estimation, observational studies, complex sample surveys, and applications of statistical methodology in law, operations management, and marketing. His recent research focuses on developing new statistical methodology with applications for real world problems.

REFERENCES

ABERNETHY, AVERY M., and GEORGE R. FRANKE. "The Information Content of Advertising: A Meta-Analysis." *Journal of Advertising* 25, 2 (1996): 1–17.

BROWN, STEVEN P., PAMELA M. HOMER, and J. JEFFREY INMAN. "A Meta-Analysis of Relationships between Ad-Evoked Feelings and

Advertising Responses." *Journal of Marketing Research* 35, 1 (1998): 114–26.

Chandy, Rajesh K., Gerard J. Tellis, Deborah J. MacInnis, and Pattana Thaivanich. "What to Say When: Advertising Appeals in Evolving Markets." *Journal of Marketing Research* 38, 4 (2001): 399–414.

Chintagunta, Pradeep K., and Naufel J. Vilcassim. "A Two-Period Repeated Game Advertising Investment Model for Oligopolistic Markets with an Application to the Beer Industry." *Decision Sciences* 26, 4 (1995): 531–59.

Dekimpe, M. G., and D. M. Hanssens. "Sustained Spending and Persistent Response: A New Look at Long-Term Marketing Profitability." *Journal of Marketing Research* 36, 4 (1999): 397–412.

Dutka, Solomon. "Combining Tests of Significance in Field Marketing Research Experiments." *Journal of Marketing Research* 31, 1 (1984): 28–43.

Erdem, Tülin, and Baohong Sun. "An Empirical Investigation of the Spillover Effects of Advertising and Sales Promotions in Umbrella Branding." *Journal of Marketing Research* 39, 4 (2002): 408–20.

Grewal, Dhruv, Sukumar Kavanoor, Edward F. Fern, Carolyn Costley, and James Barnes. "Comparative Versus Noncomparative Advertising: A Meta-Analysis." *Journal of Marketing* 61, 4 (1997): 1–15.

Hochberg, Yosef, and Ajit Tamhane. *Multiple Comparison Procedures*. New York: John Wiley and Sons, 1987.

Hollander, Myles, and Douglas A. Wolfe. *Nonparametric Statistical Methods, Second Edition*. New York: John Wiley and Sons, 1999.

Jedidi, Kamel, Carl F. Mela, and Sunil Gupta. "Managing Advertising and Promotion for Long-Run Profitability." *Marketing Science* 18, 1 (1999): 1–22.

Kaul, Anil, and Dick R. Wittink. "Empirical Generalizations about the Impact of Advertising on Price Sensitivity and Price." *Marketing Science* 14, 3 (1995): G151–60.

Leonard, Devin. "Nightmare on Madison Avenue." *Fortune*, June 29, 2004.

Lodish, Lodish M., Magid M. Abraham, S. Kalmenson, J. Livesberger, B. Lubetkin, B. Richardson, and M. E. Stevens. "How T.V. Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments." *Journal of Marketing Research* 32, 2 (1995a): 125–39.

———, ———, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens. "A Summary of Fifty-Five In-Market Experimental Estimates of the Long-Term Effect of TV Advertising." *Marketing Science* 14, 3, Part 2 (1995b): G133–40.

MacInnis, D. J., A. G. Rao, and A. M. Weiss. "Assessing When Increased Media Weight of Real-World Advertisements Helps Sales." *Journal of Marketing Research* 39, 4 (2002): 391–407.

Mela, C. F., S. Gupta, and D. R. Lehmann. "The Long Term Impact of Promotion and Advertising on Consumer Brand Choice." *Journal of Marketing Research* 34, 2 (1997): 248–61.

Naik, P. A., M. K. Mantrala, and A. G. Sawyer. "Planning Media Schedules in the Presence of Dynamic Advertising Quality." *Marketing Science* 17, 3 (1998): 214–25.

Pearson, E. S., and H. O. Hartley, eds. *Biometrika Tables for Statisticians*. Cambridge: Cambridge University Press, 1976.

Steenkamp, J. B. E. M., V. R. Nijs, D. M. Hanssens, and M. G. Dekimpe. "Competitive Reactions to Advertising and Promotion Attacks." *Marketing Science* 24, 1 (2005): 35–54.

Tellis, G. J., R. K. Chandy, and P. Thaivanich. "Which Ad Works, When, Where, and How Often? Modeling the Effects of Direct Television Advertising." *Journal of Marketing Research* 37, 1 (2000): 32–46.

Vakratsas, Demetrios, and Tim Ambler. "How Advertising Works: What Do We Really Know?" *Journal of Marketing* 63, 1 (1999): 26–43.

———, Fred M. Feinberg, Frank M. Bass, and Gurumurthy Kalyanaram. "The Shape of Advertising Response Functions Revisited: A Model of Dynamic Probabilistic Thresholds." *Marketing Science* 23, 1 (2004): 109–19.

APPENDIX

Comparison of IRI TV Advertising Test Procedures (BSCAN versus MMT)

The most fundamental difference between BehaviorScan® (BSCAN) and Matched-Market Tests (MMT) is that the former takes place at the consumer level, while the latter takes place at the store level. (The data are analyzed by IRI at the store and week level for both types of tests.)

For BSCAN testing, consumer purchasing information is electronically collected from each representative household that has been recruited and maintained in each BSCAN market. The current BSCAN markets include: Cedar Rapids (IA), Eau Claire (WI), Midland (TX), and Pittsfield (MA). The markets are large enough to test in, but small enough to allow precise control of testing variables and product distribution. A static sample of 1800–3000 households are kept in each market. Panel members shop with an ID card, which is presented at the checkouts in scanner-equipped grocery and drug stores, allowing IRI to electronically track over time each household's purchasing record at the SKU (stock keeping unit) level. For tests of alternative TV advertising plans, the BSCAN household panels are split into two or more subgroups that are balanced on past purchasing behavior, demographics, and stores that are frequented over a 1-year base period. This matching procedure makes it easier to attribute differences in the test period to the effect of the treatment rather than to preexisting differences between groups or to an interaction between the test variable and the preexisting differences. After the test ends (usually a year), the test and the results are analyzed using analysis of covariance (ANCOVA) to remove any consistent influence of all the uncontrolled factors (e.g., competitive and test brand promotions), so that the impact of the test treatment can be isolated. The split cable technology allows commercials to be substituted at the individual household level, so that one subgroup views a heavier advertising weight than the other. The data are treated as a posttest/pretest with control group design. The scanner panel purchases are aggregated to the store week level for both the treatment and control groups. A key source of sales variation unrelated to the treatments are temporary promotions, such as features, displays, price reductions, and regular price reductions, which vary across stores during the same week. To accommodate for this source of variation, a covariate is used—simply the market share for the brand in question for the total store for the week (the variable was first used by Gerald Eskin). This variable accounts for store changes at the store week level that are associated with the treatment effect. This covariate alone typically explains 20–40 percent of the variance in store-week-treatment sales over the 1-year experiment. The other covariates in the analysis are the same that were used to originally balance the experimental groups. These variables account for the variation that occurs because the different panelists will have different shopping patterns and may not stay matched for each week for the next year. The actual variables used depend on the product and category, but are straightforward demographics and past purchase variables. In all results described, the sales volume and share results are covariate adjusted using the above procedure.

MMT testing utilizes store-level scanner sales data (as opposed to BSCAN, which takes the analysis down to the household level). IRI has 64 markets with grocery scanner capabilities [e.g., Buffalo/Rochester (NY), San Diego (CA), Phoenix/Tucson (AZ), Houston (TX), and Knoxville (TN)]. Approximately half of these markets can be used for MMT tests. The MMT tests are composed of two key procedures: market matching and testing. In matching the markets, IRI considers both qualitative and quantitative factors. Qualitatively, retail environment, weather, manufacturer's sales territories, and manufacturer's national as well as regional marketing plans are considered. Markets are avoided in which other tests are being conducted or have recently been implemented that may have impact on this test. Quantitatively, IRI uses a statistical program to select the most representative test markets and the best matched control markets, based on historical data (e.g., weekly sales of the test brand and corresponding category, brand market share, etc.). The historical data used for matching covers a period of 52 weeks. In the testing stage, IRI's client executes a test treatment in the designated test markets; the control markets do not receive any treatment. At the end of the test period, ANCOVA is applied to isolate and measure the impact of the treatment that is being tested on sales. This ensures that the reported change in sales is due to the test program and not some confounding factor (e.g., test brand promotion or distribution, competitive pricing and merchandising, etc.). The purpose of the ANCOVA used in MMT tests is very similar to that of the BSCAN tests, as described earlier. The typical covariates used in the ANCOVA consist of: merchandising covariates (share of base sales volume on feature, display, and/or price reduction), price covariates (average base price per unit, or a combination of base price per unit and average percentage off base price when on deal), and other covariates, including variety (average number of UPC's weighted by dollar value), category dollar sales, and dummy variables for weeks out-of-design.

MMT has the following advantages over BSCAN: (1) it requires less in-market

execution resources from the supplier; (2) test variables can be broadly executed across a market without limiting the size of the panel; (3) it does not require drug and grocery stores as the source of the majority of the test product's sales; (4) it covers many more markets than BSCAN; and (5) for low-incidence test products, MMT's large sample size is advantageous. On the other hand, because there is less design control for an MMT test, statistical control (market matching and ANCOVA) becomes more crucial, and the validity of tests results may not be as high as BSCAN.