LEONARD M. LODISH, MAGID ABRAHAM, STUART KALMENSON, JEANNE LIVELSBERGER, BETH LUBETKIN, BRUCE RICHARDSON, and MARY ELLEN STEVENS*

The authors analyze results of 389 BEHAVIORSCAN® matched household, consumer panel, split cable, real world T.V. advertising weight, and copy tests. Additionally, study sponsors—packaged goods advertisers, T.V. networks, and advertising agencies—filled out questionnaires on 140 of the tests, which could test common beliefs about how T.V. advertising works, to evaluate strategic, media, and copy variables unavailable from the BEHAVIORSCAN® results. Although some of the variables did indeed identify T.V. advertising that positively affected sales, many of the variables *did not* differentiate among the sales effects of different advertising treatments. For example, increasing advertising budgets in relation to competitors does not increase sales in general. However, changing brand, copy, and media strategy in categories with many purchase occasions in which in-store merchandising is low increases the likelihood of T.V. advertising positively affecting sales. The authors' data do not show a strong relationship between standard recall and persuasion copy test measures and sales effectiveness. The data also suggest different variable formulations for choice and market response models that include advertising.

How T.V. Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments

We describe a meta-analytical study that tests a number of common conceptions about television advertising for consumer packaged goods. The primary data base for the study was 389 BEHAVIORSCAN® matched household, consumer panel, split cable, real world tests that were completed between June 1982 and December 1988. The only difference between the matched panels is their exposure to different T.V. advertising treatments. Sales results of the various treatments are available unobtrusively from individual level household scanner data for the matched panels.

We attempt to explain differences in the results of these advertising tests by investigating four broad classes of possible explanatory factors: (1) brand and category conditions, (2) business strategies/objectives underlying the advertising, (3) media usage (including advertising weight), and (4) copy related measures. In addition, we examine recall and persuasion measures, as provided by copy testing services, for validity in predicting the sales impact of the T.V. advertising that was tested.

The BEHAVIORSCAN® data base consists of two types of advertising tests: (1) copy tests, consisting of two test cells with different copy and equal weight (GRPs), and (2) weight tests, consisting of two cells having the same copy but different levels of exposure. We examine these two types of tests separately in the following analysis. We also analyse new products and established products separately, because the role of advertising in each situation is considered to be quite different, which leads us to the examination of four

*Leonard Lodish is the Samuel R. Harrell Professor, The Wharton School, University of Pennsylvania. Stuart Kalmenson is a doctoral student at the Sloan School, Massachusetts Institute of Technology. Magid Abraham, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens are researchers at Information Resources, Inc. (IRI), Chicago. The authors acknowledge help from David Schmittlein, Al Silk, John Farley, Paul Green, Pete Fader, and Chitrabhanu Bhattacharya. They thank the charter sponsors for their support and commitment: Leo Burnett; Campbell Soup Company; Colgate-Palmolive Company; CPC-Best Foods; DDB Needham; Dow Brands; Frito-Lay, Inc.; General Mills; Kraft General Foods; Marion Merrell Dow; Nestlé Foods Corp.; Network Television Association—ABC, CBS, NBC; Noble & Associates; NutraSweet; Pepsi Cola Company; Quaker Oats; Ralston Purina Company; Scott Paper Company; J. Walter Thompson; and two manufacturers who wish to remain anonymous. They thank the *JMR* editor and four *JMR* reviewers for their helpful comments and suggestions.

Journal of Marketing Research
Vol. XXXII (May 1995), 125–139

separate classifications of advertising tests, in addition to the examination of copy test validity.

The article is organized as follows: we discuss the study background and relevant prior research, then we describe our methodology and data, the results for each of the four classes of advertising test, and the results of the copy test validity. Finally, we discuss the implications for both advertising researchers and managers.

BACKGROUND AND RELATION TO PRIOR RESEARCH

The impetus for this meta-analytical study was a summary of the aggregate BEHAVIORSCAN® test results found in Abraham and Lodish's (1990) study. The aggregate data then analyzed showed that 49% of the electronically controlled tests of increased advertising weight had any pair of treatment advertising weights within a given test that resulted in statistically significant sales increases at the 80% level. The obvious challenge of such findings was to understand why some advertising treatments caused sales impact and others did not.

A consortium, consisting primarily of marketing research executives from the sponsor companies, was formed to collectively develop testable hypotheses and associated variables to explain the advertising test results. The hypotheses generated represent the hypotheses each company used in making advertising decisions, and most of the hypotheses are consistent with research summaries in texts such as Aaker and Myers (1988). For example, the hypothesis that advertising sales impact was related to the amount of advertising spending, particularly in relation to competition, is consistent with academic articles that posit a monotonically non-decreasing advertising budget response function. The consortium identified a number of independent variables to test these hypotheses, therefore, the BEHAVIORSCAN® data base was supplemented with questionnaire responses to capture the variables.

Most empirical research on the sales effects of T.V. advertising is based on a single brand, using primarily econometric analyses (see Aaker and Carmen 1982 or Assmus, Farley, and Lehman 1984 for summaries of the literature). An exception is Eastlack and Rao (1989) who summarize the results of 19 real environment matched market controlled advertising experiments. In the same spirit, our study attempts to find after one year cross-sectional variation in experimental advertising sales effects in a real environment of television advertising over hundreds of different brands in most of the major categories that are scanned in United States supermarkets. Another objective of our work was to replicate Blair's (1987) study of copy test validity, which showed that for established products, persuasion scores from copy tests would almost perfectly predict whether a weight (budget) increase for that copy would be significant for BEHAVIORSCAN® tests. For our replication, we were able to obtain 21 persuasion score measures for established product weight tests. Seventeen of these measures were from the same copy testing service referred to in Blair's article.

META-ANALYSIS METHODOLOGY AND THE ASSOCIATED DATA

Our analysis of the data followed the paradigm of Farley and Lehman (1986), which contains six steps. Because all of the observations in the data base derived from the same experimental environment, the analysis was easier to perform than if the observations were from different environments. Before we describe each step in our meta-analysis, we show how we treated the 40% of our tests which were multi-cell experiments. In order not to give excessive analysis weight to multi-cell experiments, in which both copy and weight were tested in the same factorially designed experiment, we randomly picked only one pair of each type from each multi-cell test to consider in the analyses. Otherwise, if we had considered all the possible pairs in complex weight-copy tests, we would not have independence between observations and would be over-weighting the advertising campaigns used in the multi-cell tests.

Step 1–Construct the dependant variables. The dependant variables used were sales volume percent changes, market share percent changes, and share point changes. The measures were adjusted to remove variation not related to the treatments, as described in the Appendix. Most previous studies used percentage gains in sales or market share associated with changes in advertising (for examples, see Assmus, Farley, and Lehman 1984). We include an absolute change measure (share point change) because absolute sales gains are related to profit, whereas percentage changes do not translate directly. Also, although it may be easier to effect percentage changes on smaller brands, absolute gains would not necessarily be related to brand size.

Step 2–Identify the independent variables. The independent variables for the meta-analysis are the inter-study differences used to specify a natural experimental design. Farley and Lehman (1986) list three types of variables to consider: (1) environmental differences, (2) differences in measurement, and (3) differences in estimation procedures. Because the same measurement and estimation procedures were used in all of the BEHAVIORSCAN® experiments, we were left to develop variables specifying the differences in research environments among the BEHAVIORSCAN® tests. In their meta-analysis of how advertising affects sales, Assmus, Farley, and Lehman (1984) specified a number of qualitative features of market environments that might affect sales results (i.e., product types and information needs, product life cycle, national setting, brand or product level, data interval, time series versus cross sectional, and media used). Given our BEHAVIORSCAN® tests were all time series, test versus control, experiments for a particular brand using television advertising over a one year duration (with sales measured weekly), for frequently purchased, relatively low-priced consumer packaged goods, most of the Assmus, Farley, and Lehman variables were constant for our data, except for the product life cycle which was included in our analysis.

In each of the four classes of explanatory factors (i.e., brand/category, strategy, media, and copy), we identified specific candidate independent variables that were selected to test the hypotheses generated by the consortium. These variables are described in Table 1.

The independent variables allow us to test a number of

hypotheses, such as whether there is a monotonically increasing sales response function to T.V. advertising (see Sasieni 1971, 1989) or whether sales response is more related to T.V. advertising budgets in relation to competitive spending (see Jones 1989).

Step 3–Analyze the natural experimental design. For the meta-analysis, our data looks like three separate matrices for each analysis cell with each cell representing a different de-

Table 1
EXPLANATORY VARIABLE LISTING

Brand and Category Variables	
* BRANDSHR	Brand Share
* BPRICE	Brand price per volume
* BRANDPEN	Brand penetration—% of households who have bought the product
* PURCYCLE	Brand purchase cycle
* BRANDREQ	Brand share of requirements—the average volume share of their purchases for triers of a brand
* BPURCH	Brand purchase occasions per buyer
* BVOLCOUP	Percent of brand volume purchased on coupon
* BVOLDISP	Percent of brand volume purchase on display
* BVOLSTORE	Percent of brand volume purchased on store coupon or feature
* CPRICE	Category price per volume
* CATPEN	Category penetration
* CATPURC	Category purchase occasions per buyer
* CATSIZE	Category size index (Category penetration × category purchase)
* PVOLCOU	Percent of category volume bought on vendor coupon
* PVOLSTOR	Percent of category volume bought on store coupon/ feature
* PVOLDISP	Percent of category volume bought on store display
* PVOLPRCE	Percent of category volume on any price deal
* RELPRICE	Brand price relative to category price
* RELPEN	Brand penetration relative to category penetration
* RELPOC	Brand purchase occasions per buyer relative to category purchases per buyer
* RELCOU	Brand volume on coupon relative to category volume on coupon
* RELDISP	Brand volume on display relative to category volume on display
* RELSTOR	Brand volume on store coupon or feature relative to category volume on store coupon or feature
PRODEST	Is the brand well established? (more than 3 years old)
REPOS	Has the brand been repositioned?
IMAGEST	Is the brand's image well established?
BRANDUN	Brand's unaided awareness level
BRANDAID	Brand's aided awareness level
* SEASON	Is the brand subject to seasonal peaks?
* IMPULSE	Is the brand likely to be purchased on impulse?
CATSTAB	Is the category stable, growing, or declining?
Strategic Variables	
What was the brand's business objective for the advertising?	
CATUSE	To increase share by increasing the penetration among current category users?
INCPEN	To increase share by increasing penetration?
NONCAT	To increase share by increasing the penetration among non-category users?
INCBUY	To increase share by increasing the buying rate?
FREQPUR	To increase share by increasing the buying rate through increased frequently of purchase?
VOLPUR	To increase share by increasing the buying rate through increased volume per purchase?
MAINFRA	To maintain the franchise?
DEFRAN	To defend the franchise?
MINSWIT	To minimize brand switching?
OBJPEN	Was the primary objective to increase share by increasing penetration?
OBJRATE	Or by increasing buying rate?
Media Variables	
MEDDOL	Dollar amount spent on media plan? (base and test)
CONT	Was the media plan continuous?
SCHED	Or flighted?
NWEEKS	The number of weeks of advertising
(PR)	Prime
(SK)	Saturday/Kid
HHGTOT	Total plan household GRPs
TAGTOT	Total plan target GRPs
DAYPADD	Did the heavy-up plan include an additional daypart (for weight tests only)
PRIMEAD	Did the heavy-up plan add prime time as a new daypart (for weight tests only)
TAREXPAN	Did the heavy-up plan include an additional target or a shift in emphasis for dual target plans (for weight tests only)?
PGRPSIN	Percent of plan GRPs delivered during introductory period (for new products only)
EXTRWGT	Was plan heavied up in front, in back, or throughout (for weight test only)?
AWKSADD	Number of extra weeks in heavy plan (for weight tests only)
BRANDLNA	What was the total amount of spending for the brand (LNA data)?
CATLNA	What was the total amount of spending for the category (LNA data)?
CTADTV	Percent of advertising in network television (Brand BAR/Brand LNA)
NONTVPR	Percent of media plan not represented by television
SOA	Share of advertising (Brand LNA/Category LNA)
SOTVA	Share of network television advertising (Brand BAR/Category BAR)
SOA/SOM	Share of advertising to share of market ratio
Copy Variables	
CREATES	What copy testing service was used?
EXEC	Were multiple executions used?
EXNO	Total number of executions
STRACH (P)	Has the copy strategy recently changed?
INTROCP	Was an introductory copy used (e.g., new flavor, new size, etc.)
MSGCHG	Was the message intended to change or to reinforce existing attitudes?
PLAYBEN	Did the advertisement play up the brand's benefits?
COMMBEN	If so, how well were these benefits communicated?
COMMSTR	Overall, how well did the creative that was executed communicate the strategy?
Copy Testing Variables	
RECALL	Category normalized recall store (Absolute Score/Category Norm)
RECNORM	Recall score relative to its norm
PERSUAS	Category normalized persuasion score (Absolute Score/Category Norm)
PERNORM	Persuasion score relative to its norm

Information for those variables with an asterisk (*) comes from IRI's BEHAVIORSCAN® database; all other information was obtained from the Sponsor Questionnaire and from IRI's BEHAVIORSCAN® Media files.

pendant variable—one for new product weight tests (N = 85), one for established product copy tests (N = 86), and one for established product weight tests (N = 207). The sample size for new product copy tests (N = 10) was too small for meaningful analysis.

Step 4 & 5–Perform the analysis & Deal with defects in the natural experiment. These steps are discussed together because the defects in the natural experiment constrain possible analysis. The number of observations for some of the variables for each analysis cell is quite small because (1) the data came from three separate sources, (2) questionnaire data were only available for 140 of the 389 tests, and (3) the questionnaires for these 140 tests were not filled in completely. The missing data, when considered with the sample sizes for each type of test, allowed multivariate analysis to be conducted only on the established product weight tests. Furthermore, for these weight tests, the increase in weight in the heavy cell was only available for 69 observations, and relatively few observations contained both weight and strategic or copy related variables. As a result of these constraints, we were unable to calibrate a stable model that included weight differences.

Because of the limitations of the data, we analyzed each of the test types by investigating summary measures and correlation matrices. We computed elasticities for the weight tests, compared them for new and established products, and discussed significant bivariate correlations. We further examined established product weight tests with multivariate analysis and identified the underlying factors with factor analysis, interpreted and used in subsequent regression analyses for each dependant variable. Finally, we examined the validity of the copy testing variables with correlation analysis.

Step 6–Draw implications. This step is covered in the subsequent sections of the article. When the reader reviews the following results, particularly for the multivariate analysis, the observations of Farley and Lehman (1986, footnote, p. 54) on the difficulties of meta-analysis should be considered: "The relative sparseness of the data makes meta-analysis in some way similar to archaeology where a few important clues and some general notions are used to construct an admittedly incomplete understanding of a situation."

RESULTS

We discuss the results for the weight tests in terms of summary measures and advertising elasticities. We then investigate in further detail the new product and established product weight tests. Finally, we consider the established brand copy tests and discuss the copy test validity.

Weight Tests – Overview

Weight tests accounted for 75% of all tests studied in this analysis. Of these, 71% were for established brands, 29% for new brands. Weight tests were classified as significant when a heavier plan outperformed, in either share or sales volume, a lower weight plan at or above the 80% one-tailed confidence level. T.V. advertising weight tests were more effective for new products, with 55% of these test pairs showing significant positive effects versus 33% for established brands. However, the magnitude of the weight increase was less for new products than established products—with the average percent increase in target GRP's being 85% for established products versus 55% for new brands. (On an absolute basis, the average level of target GRP's for established brand heavy-up plans was 2647, compared to 2520 for new brands.)

Table 2
PERCENT INCREASE IN TARGET AUDIENCE GRPs
(WEIGHT TESTS)

		Percent increase	
	N	*New Brands (85)*	*Established Brands (210)*
All Tests	141	+55%	+85%
Tests with Advertising Significant	56	+55%	+77%
Tests with T.V. Advertising N.S.	85	+54%	+88%

When the test showed a significant effect for a new product, the percent increase averaged 22% in volume and 20% in share. For successful established brand weight test pairs, the average increase was 23% for volume and 19% for share.

There were 62 established product test pairs in which one cell was no advertising, 36% of which showed a significant effect of T.V. advertising, no different from other established product weight tests.

The previous summary analysis was conducted at the 80% confidence level, thus, by chance alone 20% of the test pairs should be significant. To test whether there is a greater-than-chance likelihood of achieving such results at other confidence levels, an analysis and test recommended by Dutka (1984) was performed. Dutka shows that if the weight changes have no effect, then

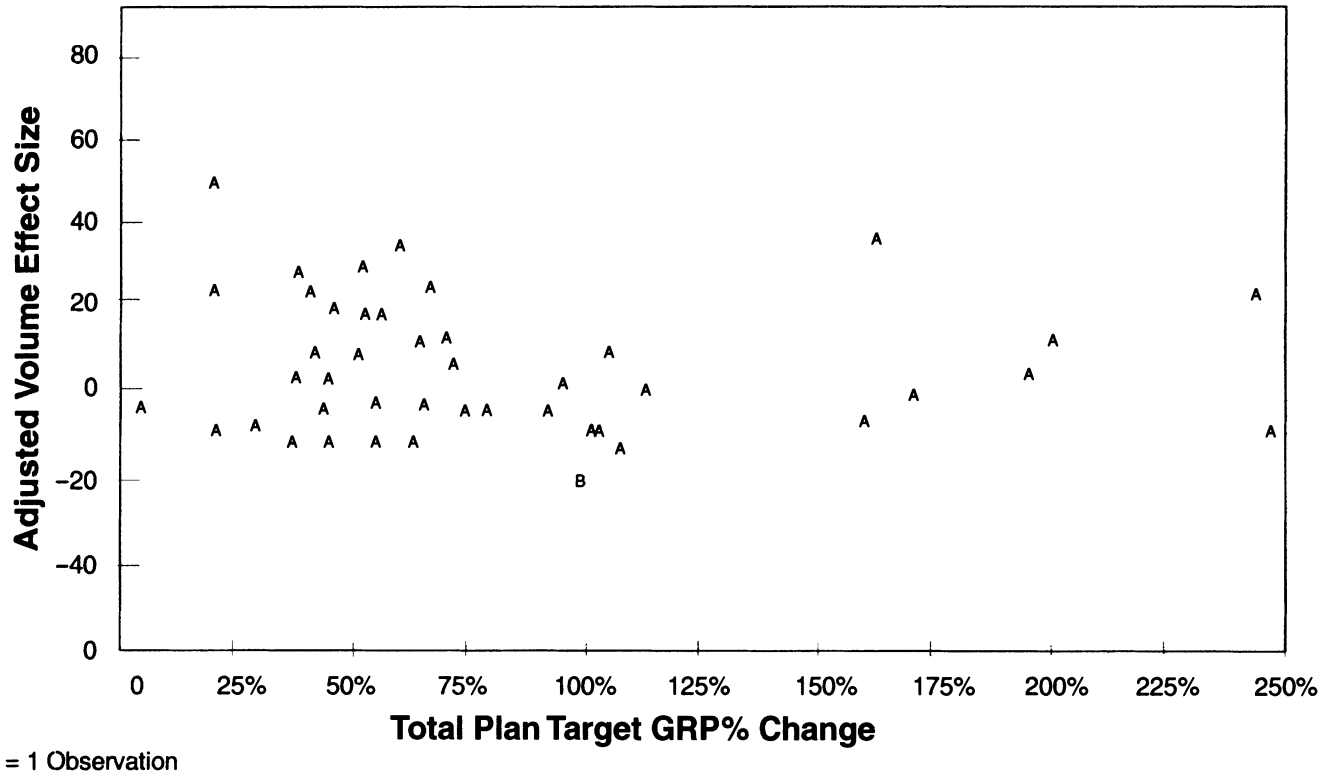
$$-2 \sum_{i-1}^{I} \ln(P_i)$$

is distributed χ^2 with 2 I degrees of freedom, in which P_i denotes the significance level for test i, i = 1 . . . I. The 75 analyzed new product weight tests had a X^2 value of 294.8 and the 210 established product weight tests had a X^2 value of 515.9, both of which were significant at less than the .005 level. Inspection of the fraction of significant tests when varying the p-value used to determine significance shows that the fraction of significant tests exceeds the p-value for all commonly used significance levels.

Weight Tests – Advertising Elasticity

Table 2 summarizes the information available for the GRP increases in weight tests. These summary findings suggest that there is no obvious relationship between the magnitude of a weight increase and the significance of the impact on sales of the advertising weight test. Figure 1 shows a scatter diagram of covariate adjusted percent volume increases versus target GRP% changes for each of the established product test pairs. Even with doubling and tripling T.V. advertising weight, there seems to be no more likelihood of achieving sales volume increases. The same type of scatter relationship is also observed for market share increases as well as any advertising weight measure including comparisons with competitive expenditures, such as share of

Figure 1
% VOLUME INCREASES ASSOCIATED WITH % TARGET GRP INCREASES FOR ESTABLISHED PRODUCT WEIGHT TESTS



T.V. advertising. As we show later, there is a tendency for larger increases in T.V. advertising weight to have a slightly negative relationship with sales increases.

For the 141 weight tests in which GRP information was available, we calculated an elasticity by dividing the estimated covariate adjusted percentage change in brand volume over the test year by the percentage change in GRP weight applied. We also separately calculated the elasticities for new versus established products. The average elasticities and standard deviations are shown in Table 3. Note that the average elasticity for established products is only significantly different from zero at the $p = .1$ level, when the elasticity for new products is five times higher and significantly less than zero at the $p < .005$ level. Although the scatter diagrams in Figure 1 show a slightly negatively-sloped sales relationship between increasing advertising and established products, the absolute average level of the points is positive.

We tested whether the difference in elasticity for new versus established products was due to the fact that the new products may be smaller in market share and therefore easier to affect percentage changes in sales.[1] We ran a regression with elasticity as the dependant variable, and two independent variables: (1) market share in the control group during the test period and (2) a dummy variable representing new versus established products. When the new and established observations are pooled, the regression shows only the new product dummy variable to be related to elasticity ($p = .01$) and the market share not to be related ($p = .89$). In terms of advertising weight elasticity, it appears that we were justified in separating new and established products for analysis.

Table 3
ADVERTISING ELASTICITIES

	N	*Average Elasticity*	*Std. Deviation*	*Std. Dev. of the Average*
All Tests	141	.13**	.40	.03
New Products	52	.26**	.49	.07
Established Products	89	.05	.32	.03

**Significant at > .01 level.

Weight Tests – Established Brands

Because there was a relatively large number of observations for established brand weight tests (N = 207), we used both correlation analysis and multivariate analysis to examine them. The correlation based analysis allowed us to include all the available data, whereas the multivariate analysis necessarily excluded variables with fewer observations but allowed for insights into interactions between the independent variables and test effects. The variables significantly correlated with any of the sales effect measures are summarized in Table 4. We found significant correlations among brand/category, strategic, media, and copy variables.

[1]We thank a JMR reviewer for this suggestion.

Table 4
SIGNIFICANT CORRELATIONS FOR ESTABLISHED BRAND WEIGHT TESTS

Variable	*Percent Volume*	*Percent Share*	*Share Point*
Brand/Category			
Brand Volume Share	**–.12 (.08) 206**	–.11 (.11) 201	–.05 (.47) 201
Brand Share of Requirements (Volume Share Among Triers)	**–.17 (.01) –201**	**–.16 (.02) 196**	–.07 (.31) 196
% Brand Volume on Display	–.10 (.25) 145	**–.14 (.08) 141**	–.02 (.80) 141
Category Growing	**.23 (.05) 72**	**.23 (.05) 72**	**.29 (.01) 72**
Brand Unaided Awareness Level	**–.25 (.04) 63**	**–.26 (.04) 63**	–.16 (.20) 63
Brand Aided Awareness Lower than Key Competitor	.21 (.12) 54	**.27 (.04) 54**	.20 (.15) 53
Brand Unaided Awareness Lower than Key Competitor	**.29 (.03) 53**	**.37 (.01) 53**	.13 (.35) 53
Brand Image Not Well Established	**.30 (.01) 92**	**.27 (.02) 72**	.19 (.11) 72
Product Well Established (3+ yrs. old)	**–.26 (.02) 74**	**–.26 (.02) 74**	–.18 (.13) 74
Strategic			
Brand/Business Objectives Included Penetration (Heavy Cell Only)	**.41 (0.0) 56**	**.33 (.01) 56**	**.46 (0.0) 56**
Media			
Heavy to Base Weight Difference: M–F Daytime Target GRPs	.21 (.15) 49	**–.27 (.07) 47**	**–.27 (.07) 47**
% M–F Daytime Target GRPs Base Cell	**–.39 (01) 49**	**–.33 (.02) 47**	–.20 (.16) 47
% M–F Daytime HH GRPs Heavy Cell	**–.27 (.08) 42**	–.21 (.19) 42	–.19 (.24) 42
Extra Weight in Heavy Cell Added to Both Front and Back End (versus Front Only/Back Only)	.01 (.88) 69	.05 (.68) 67	**–.23 (.05) 67**
Heavy to Base Weight Difference: Media Dollars Spent	–.21 (.11) 59	**–.29 (.03) 58**	–.19 (.15) 58
Heavy to Base Brand Share of Network T.V. Advertising: Percentage Point Difference	–.28 (.18) 25	**–.36 (.07) 25**	**–.49 (.01) 25**
% of Brand Advertising Dollars Spent on Network T.V. (BAR/LNA)	–.10 (.36) 85	**–.18 (.09) 85**	.07 (.52) 85
Brand Share of Network T.V. Advertising (SOTVA)	**–.23 (.01) 110**	**–.23 (.02) 110**	–.101 (.31) 110
Target Audience Expanded/Emphasis Shift	.09 (.16) 217	.08 (.24) 208	**.132 (.06) 208**
Primary Target Audience Teens	–.09 (.29) 135	–.09 (.28) 132	**–.18 (.04) 132**
Copy			
Copy Message Intended to Reinforce Attitudes	**–.26 (.03) 66**	**–.24 (.05) 66**	**–.25 (.05) 66**
Copy Strategy had not Changed Recently	–.13 (.31) 67	–.14 (.25) 67	**–.33 (.01) 67**
Copy was *not* Introductory	–.17 (.17) 66	**–.21 (.09) 66**	–10 (.41) 66

The data implies that it is easier for less well-established and smaller brands than for well-established brands to effect a change with increased weight. The results are not surprising in light of what was seen in the analysis of advertising elasticity. As new products were more responsive to increased weight than established products, so were less entrenched brands more responsive among the established products. Two other variables that positively correlated with test success on a percentage change basis were (1) lower than average aided and unaided awareness measures and (2) brand images that were not well established. A growing category was correlated positively with test success regardless of the dependant variable used to measure test effectiveness.

Only a single strategic variable, brands whose business objective included increasing penetration, was correlated with larger test effects; however, other strategic variables had significant coefficients in the multivariate analysis. Such findings indicate that care must be taken in the interpretation of bivariate results and that interactions among independent variables may support the significance between the bivariate correlations and the test effect variables.

Such interactions may serve as explanations for many of the significant media variables included in Table 4. There were no significant positive relationships exhibited between the percent increase in GRPs and percent sales increases. In fact there were some significant negative correlations between advertising weight and test effects, which were found with absolute dollar differences, percentage point differences in brand share of television advertising, absolute national share of television advertising, and the percentage of advertising expenditures allocated to television. This may be explained, however, by interactions between advertising expenditure and brand size. Larger absolute weight differences are associated with heavier spending brands, which tend to be the larger—and better—established brands. Brand size (share) was shown to be negatively correlated with test effect.

Targeting also appears to play a role in test effects. Tests in which the heavy plans included either an expansion or a shift in emphasis of the target audience definition (versus the base plan) were correlated with larger test effect sizes.

Weight Tests – Established Brands – Multivariate

The first steps in the multivariate analysis were the handling of missing data and the manipulation of the raw data set. The variables for which a large majority of data was missing, the variables with variances very near zero, and the dummy variables with less than 15 observations of zero or one were removed from consideration. The remaining variables were subject to exploratory factor analysis.

Visual inspection of the sponsor questionnaires showed that they were filled out in clusters or groups of variables (e.g., all media, strategic, brand/category or copy variables would be missing on the same questionnaire). As a result, a factor analysis of the total independent variable data set was unstable and impossible to interpret. Consequently, separate factor analyses were conducted for each category of vari-

ables, using listwise deletion of cases with missing values. The rotated factor matrices for the brand/category, and strategic classes of variables, as well as our interpretation of the labeling of each factor are reported in Tables 5 and 6.

The factor analysis for the brand/category and strategic variables represented 91.6% and 91.5% of the variance, respectively. We computed factor scores using the regression method. Although the copy and media variables were best represented by a set of dummy variables because of sample size constraints and unstable factor analyses, we used as candidates any media and copy variables that had enough observations to have reliable bivariate correlations with the other independent and dependent variables. The correlations of factor scores and the dependent variables are included in Table 7 for completeness.

For these established product weight tests, the increase in weight in the heavy cell was only available for 69 observations, which proved problematic in further analysis. A number of models including the weight were devised, but the missing data prevented the estimation of the models. Because of the missing values, no statistically significant results were found in attempting to calibrate the models that included the weight difference. There are also relatively few observations with both weight and strategic variables, or with both weight and dummy variables.

Regression models were estimated for the previous predictor variables with the three dependent variables: covariate adjusted volume effect size, covariate adjusted share effect size, and share point change. We incorporated the incomplete observations by using all valid pairs of data to compute the correlation matrix. The correlation computed with the fewest number of observations (41) was used as the sample size to conservatively compute regression statistics. We then eliminated the predictors that were not significant in any of the three models and reestimated the parameters. The results of the regressions are summarized in Table 8.

In the spirit of the Farley and Lehman (1986), these multivariate results must be interpreted quite conservatively, because there are not enough observations to do a split half validation. We were able to randomly eliminate 15% of the data points from each of the three regressions for three replications. In all of these replications with 15% fewer points, none of the significant coefficients changed signs or moved more than their 95% confidence limits.

A number of variables seem to explain sales impact in the multivariate analysis. Advertising messages that are intended to change, not reinforce, existing attitudes (MSGCHP)

Table 5

ROTATED FACTOR MATRIX AND INTERPRETATION BRAND AND CATEGORY VARIABLES

	SCR 1 Factor 1 (category size)	*SCR 2 Factor 2 (brand share)*	*SCR 3 Factor 3 (couponing)*	*SCR 4 Factor 4 (price)*	*SCR 5 Factor 5 (display)*	*SCR 6 Factor 6 (price reduction, feature)*	*SCR 7 Factor 7 (relative featuring, price reduction*	*SCR 8 Factor 8 (relative couponing)*	*SCR 9 Factor 9 (purchase cycle)*	*SCR 10 Factor 10 (relative display)*	*SCR 11 Factor 11 (relative price)*
CATSIZE	**.94**	.08	–.04	.03	.11	.07	–.01	.00	.01	–.01	.00
CATPURC	**.94**	.08	–.02	.04	.04	.03	.01	–.02	–.05	–.04	–.01
RELPOC	**–.80**	.33	.01	.10	–.08	–.06	–.16	–.09	–.14	.17	.07
CATPEN	**.79**	.07	–.06	–.04	.21	.24	.02	.01	.04	.02	.00
BRANDREQ	**–.73**	.57	–.08	.05	–.03	–.08	.00	–.13	.09	.05	.01
RELPEN	–.04	**.94**	–.12	.00	.07	.11	.08	–.01	.06	–.04	.00
BRANDSHR	–.39	**.87**	–.05	.00	.00	–.03	.07	–.08	.08	.00	–.03
BRANDPEN	.36	**.79**	–.16	–.02	.18	.26	.09	.00	.03	.00	.00
BPURCH	.45	**.61**	–.05	.17	–.06	.09	–.13	–.16	–.38	.05	.04
PVOLCOU	.05	–.10	**.89**	.12	–.10	–.06	–.14	–.24	.08	.05	–.03
BVOLCOUP	–.07	–.22	**.87**	.13	–.11	–.10	–.08	.22	.01	–.02	–.02
CPRICE	–.03	–.01	.05	**.94**	.00	.02	–.08	.04	–.02	.01	–.09
BPRICE	.01	.04	.18	**.92**	–.02	.02	–.07	–.01	.04	–.06	.03
BVOLDISP	.12	.08	–.09	–.04	**.90**	.02	.01	–.03	.01	.22	–.12
PVOLDDISP	.28	.06	–.05	.00	**.80**	.25	–.15	–.05	–.03	–.21	.14
PVOLSTOR	.30	.20	–.16	.06	.18	**.82**	–.02	.05	–.01	–.12	.00
PVOLPRCE	.12	.02	.56	.09	.29	**.61**	.11	–.19	.01	–.15	.04
RELSTOR	.04	.07	–.08	–.12	–.04	–.03	**.94**	.00	–.04	.14	.04
BVOLSTOR	.10	.14	–.19	–.09	–.10	.60	**.67**	.01	–.18	–.02	
RELCOU	.07	–.11	–.05	.02	–.06	.00	.00	**.97**	.00	–.03	–.03 –.04
PURCYCLE	.05	.09	.07	.03	–.01	–.02	–.08	–.01	**.95**	.02	.06
RELDISP	–.11	–.02	.00	–.04	.05	–.13	.12	–.03	.02	**.94**	.05
RELPRICE	–.04	–.01	–.03	–.06	–.01	.00	–.04	.05	.04	–.11	**–.97**

Table 6
ROTATED FACTOR MATRIX AND INTERPRETATION: STRATEGIC VARIABLES

	Strat 1 Factor 1 (increasing buying rate)	*Strat 2 Factor 2 (maintain franchise)*	*Strat 3 Factor 3 (increase penetration)*	*Strat 4 Factor 4 (objective penetration)*	*Strat 5 Factor 5 (increase volume purchase)*	*Strat 6 Factor 6 (non-category users)*	*Strat 7 Factor 7 (defend franchise)*
OBJRATEP	**.83**	–.26	.00	–.33	.11	.06	–.10
FREQPURP	**.78**	–.06	–.03	.33	.28	–.10	–.11
INCBUYP	**.69**	–.37	.18	.36	.08	–.13	–.12
MAINFRAP	–.16	**.91**	.08	–.09	.03	.05	.17
MINSWITP	.26	**.78**	.25	–.08	–.17	–.04	.23
INCPENP	–.02	.09	**.96**	.12	.00	.05	.00
CATUSEP	.15	.31	**.72**	–.04	.04	.09	.07
OBJPENP	.03	.13	.09	**.95**	–.02	.02	–.01
VOLPURP	.21	–.06	.02	–.02	**.96**	–.02	–.05
NONCATP	–.04	.01	.07	.01	–.02	**.99**	.03
DEFRANP	–.16	.32	.02	–.01	–.06	.04	**.92**

and changes in copy strategy (STRACHP) seem to positively affect the sales impact of advertising weight changes. Advertising and media plans with objectives of both increasing penetration to new users and increasing, rather than maintaining, buying rates are also positively related to sales changes. The data, however, point to the mitigation of these effects by introducing INTROCPP, small changes to the product offering (e.g., new packaging, sizes, formulations, or flavors). Although INTROCPP was not significantly related to any of the three sales effect measures on a bivariate basis, when the other variables that involved more fundamental changes in advertising were added, INTROCPP became significantly negative. These data support the hypothesis that introducing small changes to the product offering may be counterproductive and produce diminishing returns.

Advertising weight changes seem to affect categories that have many purchase occasions and a high penetration the easiest. Brands with smaller shares amd lower prices, and brands that are less than three years old also seem relatively easy to influence on an absolute as well as percentage change basis. In store displays, both absolute and relative, show a negative relationship with test effect size, which may be because the effects of the in-store display tend to overshadow the differences in advertising weight.

The standard flighted media plan, though not related to sales effects on a bivariate basis, becomes significantly negatively related in the regression. The hypothesis that changes in the media plan—with big bumps rather than periodic flights—may work better (Strong 1977) and is supported further by the new product weight tests mentioned subsequently.

Table 8 indicates that the measure of sales effect, percentage changes or absolute changes, does not have a major effect on the variables that seem to be associated with it. The coefficients are very similar across the three regression models. Although this may be an artifact due to the similarity of the subsamples that had bivariate observations used for the three underlying correlation matrices, our limited examination of hold-out samples does not support this reasoning.

Weight Tests – New Brands

For new products, many of the variables included on Table 1 were not relevant. For example, brand characteristics and relative spending levels were yet to be established. Many strategy and copy variables were also not applicable. Category characteristics and media variables had significant correlations with many of the dependant variables for new product weight tests and are summarized in Table 9.

In terms of category characteristics, Table 9 indicates significant positive correlations found with test sales effects for brands in larger categories, in which the number of category purchase occasions were higher and the variety of products from which to choose was greater. This suggests that as category size increases and overall category purchase occasions increase, a new player will have more opportunities to capture sales, and advertising will be more likely to have an impact on switching behavior.

The percentage of category volume bought on manufacturer coupons correlated positively with increased volume sales but not with share. This suggests that there can be a synergistic effect of advertising weight for new brands on heavier coupon users in which advertising causes a heightened awareness of new brands, an increase in clipping product coupons when seen, and a possible expansion of the market. At the same time, exposure to the coupon provides additional impressions on the consumer.

There also tended to be higher test effects for increased weight efforts behind new brands in "impulse item" categories, which suggests that advertising may play an important role in reminding consumers to pick up those new items that are generally consumed quickly upon availability.

Media related variables also showed significant correlations with test effects. Specifically, larger differences in prime time target GRPs were correlated with increased sales. Prime T.V. is the largest reach daypart. This finding is consistent with the media plans targeting new people.

In terms of scheduling incremental weight, we found a negative correlation between test effects and adding weeks

to the plan. This suggests that the concentration of advertising weight may offer advantages over its dispersion, and illustrates the possible importance of overcoming noise by advertising at significantly higher weight levels when advertising for a new product. In the same way, it seems important to stimulate trial early on with heavy introductory weights. The heavy cell plans that placed the emphasis on the introductory period relative to the sustaining period had higher test effects than those that were proportionately less front-loaded.

Copy Tests – Overview

We analyzed a total of 96 copy test pairs, 90% of which were for established brands. Copy tests were considered significant when the copy airing in either cell outperformed the other at the 80% (two-tailed) confidence level. For new brands, 60% of the tests showed significant effects, versus 25% for established brands. It should be noted, however, that only 10 new brand copy tests were examined. Significant new brand copy test sales effects showed an average effect of 23% on volume and 19% on share, versus the effects of 23% on volume and 12% on share for established brands.

To test whether there is a greater-than-chance probability of achieving these results, we repeated the Dutka (1984) procedure. For established brands, the χ^2 test for the 84 pairs is 155.5, which is significant at the .005 level. However, visual inspection of the number of significant tests at different significance levels, indicates that often the p-value exceeds the fraction of significant tests. This indicates that chance in fact may account for most of the copy test sales effect variation. We have reason to believe that the test pairs represent minor executional subtleties that would be unlikely to have a meaningful impact. The copy pairs do not show changes on many of the copy-related variables that seemed to help increased weight to work.

Table 4 includes a number of copy related variables that showed significant correlations with test effects for established product weight tests. Test effects tended to be higher when the message was intended to change attitudes rather than reinforce them, when the copy strategy had recently changed, and when copy was introductory in nature. Copy was defined as introductory for established brands when the message included introduction of new elements such as flavor, brand form, reformulation, and packaging. Evidence here supports the importance of keeping the message fresh.

Keeping in mind the above correlation results for weight tests, it is interesting to summarize the copy and strategy differences tested for the established product copy tests. Based on questionnaire responses, only 9.1% of the copy tested pairs differed in their objective between changing or reinforcing existing attitudes; only 2.7% tested introductory versus non-introductory copy; and only 16.7% had one of the pairs testing an altered copy strategy. In terms of brand strategy, only 9.7% of the tested pairs had differences in whether the brand strategy was to increase penetration. Thus, there were very few tested copy pairs that were different on the copy/brand variables that seemed to effect weight response. We contend that in the copy tests discussed subsequently few real, substantive copy differences were tested in our data base for established product copy tests.

Table 7

BIVARIATE CORRELATIONS WITH DEPENDENT VARIABLES (N)
P VALUE

Brand and Category Variables	*Percentage Volume*	*Percentage Share*	*Share Point*
Category Size	.084	.061	–.026
	(137)	(133)	(133)
	P = .329	P = .486	P = .765
Brand Share	–.141	–.155	–.123
	(137)	(133)	(133)
	P = 0.93	P = .073	P = .157
Couponing	–.022	–.003	–.099
	(137)	(133)	(133)
	P = .796	P = .972	P = .253
Price	–.058	–.056	–.095
	(137)	(133)	(133)
	P = .500)	P = .518	P = .274
Display	–.142	–.202	–.138
	(137)	(133)	(133)
	P = .096	P = .020	P = 113
Price Reduction, Feature	–.023	–.028	.012
	(137)	(133)	(133)
	P = .781	P = 741	P = .887
Relative Featuring, Price Reduction	–.051	–.057	–.081
	(137)	(133)	(133)
	P = .550	P = .508	P = .353
Relative Couponing	.083	.072	.085
	(137)	(133)	(133)
	P = .331	P = 409	P = .328
Purchase Cycle	.084	.059	–.085
	(137)	(133)	(133)
	P = .329	P = 494	P = .326
Relative Display	–.013	–.001	.089
	(137)	(133)	(133)
	P = .872	P = 988	P = .308
Relative Price	–.000	–.011	.140
	(137)	(133)	(133)
	P = .999	P = 895	P = .107
Strategic Variables	*Percentage Volume*	*Percentage Share*	*Share Point*
Increasing Buying Rate	.089	.040	.125
	(72)	(72)	(72)
	P = .454)	P = 734	P = .296
Maintain Franchise	–.064	–.042	–.125
	(72)	(72)	(72)
	P = .588	P = .720	P = .295
Increase Penetration	.105	.003	.002
	(72)	(72)	(72)
	P = .379	P = .976	P = .983
Objective Penetration	.019	.032	–.029
	(72)	(72)	(72)
	P = .872	P = 788	P = .809
Increase Volume Purchase	–.046	–.029	–.016
	(72)	(72)	(72)
	P = .699	P = .803	P = 891
Non-Category Users	.004	–.021	.105
	(72)	(72)	(72)
	P = .970	P = 855	P = .378
Defend Franchise	–.091	–.068	–.119
	(72)	(72)	(72)
	P = .444	P = .570	P = .316

Table 8
ESTABLISHED BRAND WEIGHT TESTS

Independent Variables	*Dependent Variable*					
	% Change in Volume		*% Change in Share*		*Share Point Change*	
	Beta	*P*	*Beta*	*P*	*Beta*	*P*
SCR 1 (Category Size)	.37	.01	.27	.102	.26	.115
SCR 2 (Brand Share)	–.46	.0008	–.38	.014	–.38	.012
SCR 4 (Price)	–.51	.0012	–.43	.018	–.42	.017
SCR 5 (Display)	–.78	.0000	–.70	.08	–.69	.005
SCR 10 (Relative Display)	–.37	.0136	–.26	.13	–.27	.105
STRAT 1 (Increase Buying Rate)	.32	.0675	.26	.21	.31	.124
STRAT 3 (Increase Penetration)	.40	.0025	.23	.12	.26	.07
STRAT 6 (Non-Category Users)	.18	.16	.103	.51	.24	.112
MSGCHP (Message Intended to Change Attitudes)	.90	.0000	.72	.003	.71	.002
STRACHP (Change in Copy Strategy)	.475	.0014	.40	.02	.57	.001
INTROCPP (Introduce New Size/Flavor)	–.57	.0012	–.38	.060	–.46	.019
PRODUCT LESS THAN 3 YRS OLD	.25	.102	.21	.25	.14	.40
SCHEDPC (Flighted Media Plan)	–1.02	.0000	–.85	.002	–.83	.001
Adjusted R^2	.471		.237		.290	
F (12/34)	4.22		2.12		2.47	
Significance of F	.0004		.039		.017	

Although the number of tests for new brands is small, the results clearly suggest that new brands tend to be more sensitive to copy differences than established brands. The Dutka procedure shows that sales effects of new copy tests exceed chance at any significance level. Because of the very small sample size for the new brand copy tests, we will present only the above aggregate summary.

Copy Tests – Established Brands

The brand and category variables that were correlated with sales effects for established brand copy tests are summarized in Table 10. Many of the relationships mirror those found for weight tests and support the assertion that percentage effect changes are harder to affect for larger brands. Note here that the dependant variable is the absolute value of the volume or share change because the direction of the change is irrelevant.

Copy Service Variables – Copy Test Validity

We next discuss the results of our examination of the relationship between sales changes and standard copy service measures of recall and persuasion for established brands. Our intent was not to evaluate one service against the other, but to look at scores in general relative to BEHAVIORSCAN® findings and replicate the analysis of Blair's (1987) study with later data. Sample sizes again limited analysis to established brand weight and copy tests and new brand weight tests. For established brand weight tests, there was sufficient information to examine both recall and persuasion scores relative to test success. For tests of one copy versus another and new brand weight tests, sample sizes permitted an evaluation of recall scores only. The sample size for established weight persuasion tests was 21, established weight recall tests 41, and established copy recall tests 16. For the weight tests there was considerable variation in the recall and persuasion scores for the copy executed in the tests. Likewise, for the copy tests, there was considerable variation in the recall score differences between the pairs of tested copies.

We examined the relationships between recall and persuasion measures and sales effects in several ways. First, because multiple syndicated services and in-house organizations were represented, correlations were based on scores indexed to their category norms (absolute scores vary and norms, against which significant differences are gauged, vary across categories, brands, and services). Second, we examined a relative ranking of above, at, or below the norm. Third, we investigated the relationship between copy test

Table 9
SIGNIFICANT CORRELATIONS FOR NEW BRAND WEIGHT TESTS

Variable	*Percent Volume*	*Percent Share*	*Share Point*
Brand/Category			
Category Purchase Occasions/Buyer	**.20 (.10) 73**	.176 (.15) 69	–.05 (.70) 69
Category Size			
(Penetration × Purchase Occasions)	**.21 (.08) 72**	.16 (.18) 69	–.05 (.71) 69
% Category Volume Bought on Vendor Coupon	**.22 (.05) 76**	–.005 (.96) 71	–.09 (.47) 71
Impulsive Purchase Category	**.20 (.09) 75**	.19 (.12) 70	.03 (.78) 70
Media			
Heavy to Base Weight Difference: Prime Time Target GRPs	**.29 (.08) 39**	**.31 (.06) 38**	.21 (.20) 38
Heavy to Base Weight Ratio: Prime Time Target GRPs	.15 (.42) 29	**.41 (.03) 28**	.30 (.11) 28
Heavy to Base Weight Difference: # of Weeks of T.V. Advertising	–.18 (.25) 41	**–.27 (.10) 39**	–.16 (.33) 39
Heavy versus Base: % GRPs Intro Period Difference	.21 (.31) 25	**.36 (.08) 24**	.05 (.79) 24

scores and the six-month sales effects to control for the possible complicating factor of commercial wear-out.

For new and established brand weight tests, there was no significant relationship between any measure of recall and sales impact. The relationship between recall differences for established brand copy test pairs and persuasion scores for established brand weight tests and their sales impact was tenuous. Specifically, although there were significant correlations, the significance of the correlations of recall scores for established brand copy tests and persuasion scores for established brand weight tests depended on one or two points. For recall, the significance (and direction) of the relationship depends on two points at the extremely high end of the difference between scores for the two copies tested; for persuasion, these points are at extremely low or extremely high levels of category normalized persuasion—suggesting the possibility that scores below or above these thresholds might be predictive of success at generating increased sales with higher weight. Between extremes there seems to be no relationship between persuasion or recall scores and sales changes. Figures 2A and 2B show the "best" relationships found for persuasion and recall.

These results should be compared with those of the Advertising Research Foundation (ARF) Copy Research Validity Project (Haley and Baldinger 1991). The ARF study used only copy tests, whereas our study on persuasion is based on weight test results. Our recall results are based on 16 different copy tests, each of which tested a pair of copies. Our results then are for 32 copies versus the 10 copies used in the ARF study. Our correlation, or scatter plot, analysis determines not only whether the technique can pick the better of two copies, but it also evaluates *how much* better (in sales) one copy was versus the other.

It should be noted that additional investigation included examination of absolute scores by service and analysis of scores excluding those obtained more than two years prior to the start of the BEHAVIORSCAN® test. We also examined multiplying recall by persuasion for the established product weight tests in which the data were available and did a multiple regression of recall, persuasion, and sales increases. The results from these analyses did not change the tenuous relationship between copy test recall and persuasion scores and sales changes.

We compare our results with those of Blair (1987), because it is a replication of her work. Table 11 shows our data for absolute persuasion scores and advertising test significance. Blair reports a perfect relationship between persuasion scores and test success, with all persuasion scores below 5.9 associated with non-successful tests and all persuasion scores at or above 5.9 associated with successful tests. Table 11 indicates that we were unable to replicate these results in our later examination of T.V. advertising tests, due to the existence of both significant tests with relatively low persuasion measures, and non-significant tests with relatively high persuasion measures. Perhaps the T.V. advertising environment with many more shorter commercials is different in our data, and more than persuasion is now needed to be sure of affecting sales.

It also should be mentioned that these data were gathered retrospectively and represent how sponsors actually used the persuasion pretests, not a subset of pretests that were screened. Some of the recall and/or persuasion tests were done on rough copy executions stated to have been similar to the actual campaigns that ran, which is a very common practice. However, the judgments of similarity to what finally ran is inherently subjective. It is not clear from Blair's (1987) study what was assumed or screened. See Blair and Rosenberg (1994) for a different point of view.

IMPLICATIONS AND CAVEATS

Implications and Caveats for Academic Researchers

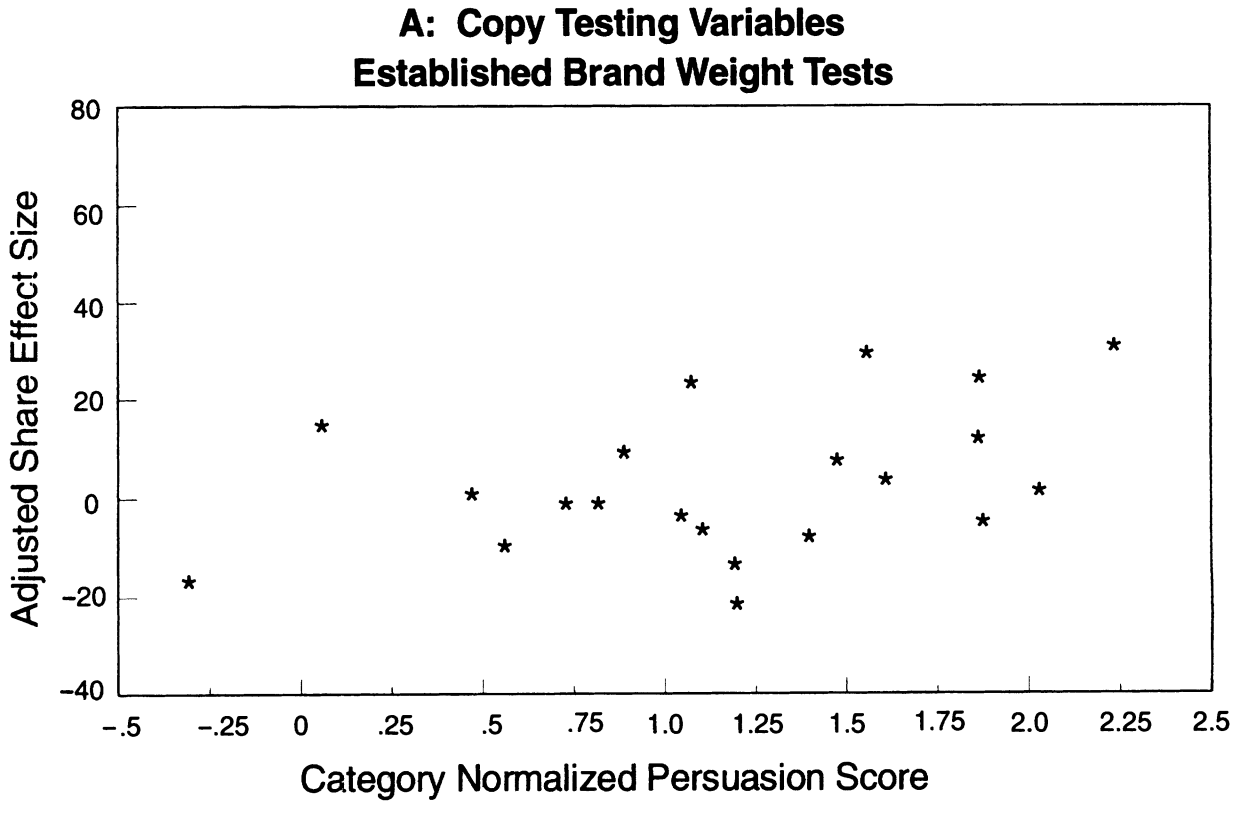
This study calls into question the assumption that there is usually a monotonically increasing T.V. advertising response function for an established brand. Given that 67% of the established brand weight tests are not significant (i.e., at the .2 level) when large amounts of advertising are added, and that there is, if anything, a negative relationship between

Table 10

SIGNIFICANT CORRELATIONS FOR ESTABLISHED BRAND COPY TESTS

Variable	*Percent Volume*	*Percent Share*	*Share Point*
Brand/Category			
Brand Purchase Occasions Per Buyer	**–.30 (.02) 59**	**–.35 (.01) 56**	**–.29 (.03) 56**
Brand Penetration	**–.47 (.00) 82**	**–.47 (.00) 78**	.08 (.43) 78
Brand Share of Requirements	–.07 (.54) 80	**–.20 (.08) 77**	**.49 (.00) 77**
Brand Volume Share	**–.22 (.04) 82**	**–.33 (.00) 78**	**.49 (.00) 78**
% Brand Volume on Vendor Coupon	.17 (.19) 59	**.32 (.02) 56**	.06 (.65) 56
% Brand Volume on Display	.17 (.18) 59	**–.29 (.03) 56**	**–.26 (.05) 56**
Category Penetration	**–.25 (.02) 86**	–.17 (.12) 82	**–.41 (.00) 82**
Category Purchase Occasions/Buyer	–.07 (.54) 85	–.09 (.42) 81	**–.20 (.08) 81**
Category Size (Penetration × Purchase Occasions)	–.17 (.12) 85	–.14 (.20) 81	**–.29 (.08) 81**
% Category Volume Bought on Store Coupon/Feature	**–.22 (.04) 86**	**–.20 (.07) 82**	–.10 (.36) 82
Brand Penetration Relative to Category Penetration	**–.41 (.00) 82**	**–.48 (.00) 78**	**.21 (.06) 78**
Brand's Aided Awareness	**–.40 (.03) 30**	**–.33 (.08) 30**	–.05 (.77) 31
Brand's Unaided Awareness	**–.40 (.03) 30**	**–.33 (.08) 30**	.06 (.75) 30
Brand Aided Awareness Lower than Key Competitor	**.34 (.07) 28**	.30 (.12) 28	–.19 (.33) 28
Brand Unaided Awareness Lower than Key Competitor	**.62 (.00) 29**	**.50 (.01) 29**	.08 (.67) 29
Media			
Brand Share of Advertising (SOA)	**–.21 (.10) 58**	**–.22 (.09) 58**	–.13 (.34) 58
Brand Share of Network T.V. (SOTVA)	–.12 (.37) 57	**–.23 (.08) 57**	–.02 (.85) 57

Figure 2
RECALL AND PERSUASION COPY TESTING SCORES VERSUS SALES IMPACT



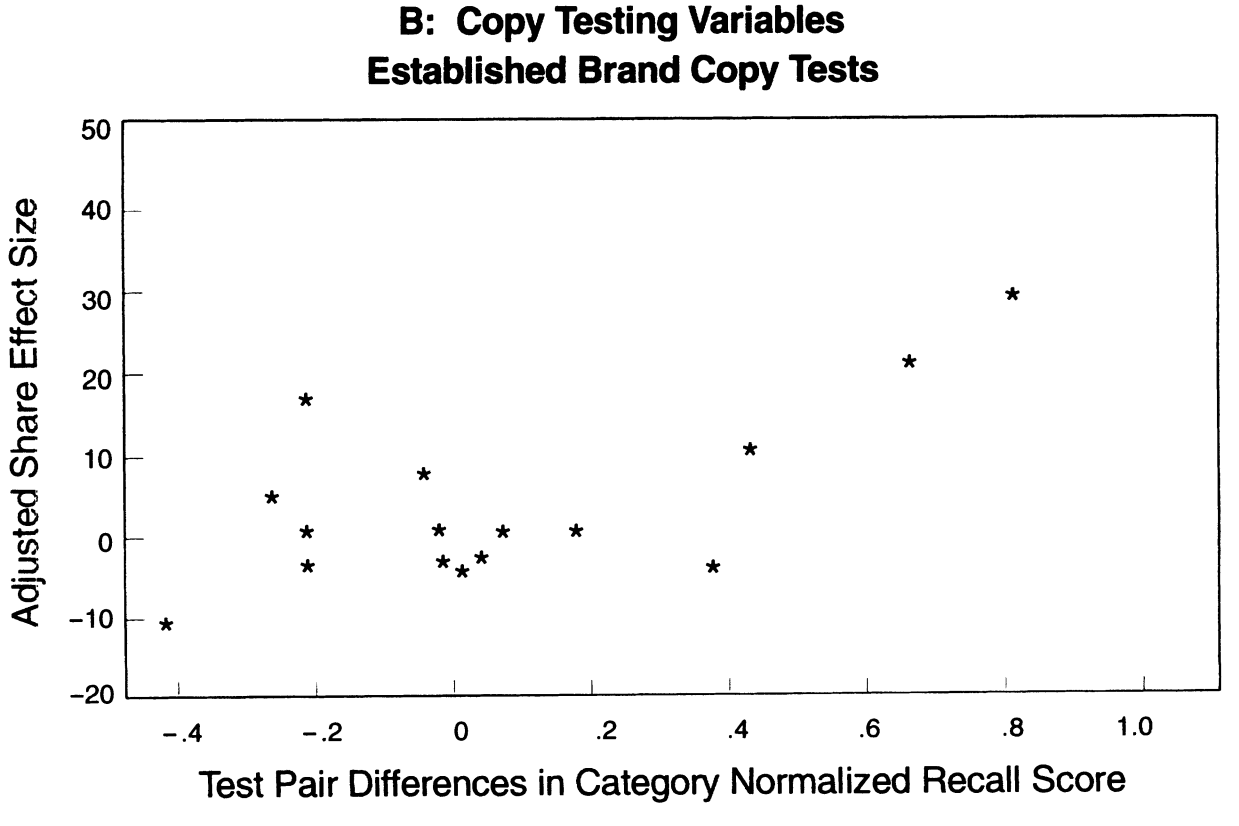


Table 11
ABSOLUTE PERSUASION SCORES FOR ESTABLISHED BRAND WEIGHT TESTS

Absolute Persuasion Score	*Covariate Adjusted Volume Confidence Level*	*Covariate Adjusted Share Confidence Level*	*Significant (80%) Sales Difference Volume or Share*	*Signficant (90%) Sales Difference Volume or Share*
−1.0	12	9	No	No
0.6	54	86	Yes	No
1.9	16	25	No	No
2.6	5	5	No	No
2.8	78	98	Yes	Yes
3.0	62	63	No	No
4.0	83	83	Yes	No
4.3	27	74	No	No
4.7	50	32	No	No
5.0	81	82	Yes	No
5.4	99	99	Yes	Yes
5.4	63	60	No	No
5.9	44	49	No	No
6.2	92	85	Yes	Yes
6.2	25	26	No	No
6.2	77	81	Yes	No
7.5	95	98	Yes	Yes

increasing T.V. advertising weight—or increasing share of voice or share of T.V. advertising—and sales volume or share changes, the articles that first specify one non-decreasing, monotonic advertising response function for a brand or brands and then develop strategies based upon their assumed response function could need to reexamine their assumptions. The only way our data could be consistent with a monotonically nondecreasing response function is if most of the brands are already at a unique saturation level, if those brands have unique response functions, and if the saturation levels for brands that have higher percentage increases in advertising have lower saturation levels. (See, for example, Jones 1989; Mahajan and Muller 1986; Sasieni 1971, 1989; Simon (1982), Simon and Arndt 1980; or Spence 1980.)

Also, given the wide dispersion of sales effects of T.V. advertising for established brands in Figure 1, the micro-level choice model parameterizations that postulate a common effectiveness function for all competing brands and copy platforms should be revisited. (See Deighton, Henderson, and Neslin 1994 for a recent example.) One of the authors has parameterized similar logit models with separate parameters for each copy platform for each competitive brand. Significantly different effectiveness coefficients are found by brand and by copy platform.

The average elasticity of advertising for established products' weight implied by this study is much lower than that found by Assmus, Farley, and Lehman (1984) for their analysis of econometric studies—about .05 versus about .15. However, the Assmus, Farley, and Lehman estimate is not clearly comparable because of all of the design variables included in their study universe. The .15 results from subtracting coefficients for annual data—not for including carryover or including exogenous variables—from their grand mean in the ANOVA for their meta-analysis. Our results are similar in magnitude to the experimental results summarized by Aaker and Carmen (1982). Perhaps there is more bias inherent in the econometric models analyzed by Assmus, Farley, and Lehman (1984) than was teased out by their meta-analysis. Our new product weight data is consistent with Rhodes (1977), who reported that new product ADTEL weight tests were much more likely to show significance than established products.

Another factor differentiating the current study from earlier studies is its sample. Our test base for weight tests may not be typical of all T.V. advertising for packaged goods because it is not a random sample. One possibility is that the brands selected for advertising tests were brands that were "in trouble." We examined the control group share changes over a one-year period, and found that approximately half showed increases in share and that the average share change in the control group was not significantly different from zero. This indicates that our sample does not consist solely of "problem brands." However, because the great majority of the tests were increases in advertising weight, it may be reasonable to assume that the advertiser assumed that the campaign would be productive if more weight were added. Thus, the bias in the sample might be toward more effective advertising, which would make our elasticity estimate optimistic.

The bivariate and multivariate relationships that we found in this empirical data should be useful for future researchers to test and diagnose. Although the multivariate relationships we developed are conceptually more appealing than the bivariate, we believe there also is value in the bivariate correlations. Many of these correlations have many more observations underlying them than the multivariate ones. As long as we realize that the variables may be correlated with other independent variables, which our factor analysis shows, then their interpretation can still provide a good summary of how various variables as they are used in practice with other variables can affect the sales impact of T.V. advertising for established products. For example, the relationships we found between T.V. advertising and different types of sales promotions should provide rich opportunity for studies to amplify and explain the phenomena. The tenuous nature of the relationship between recall and persuasion with sales effect in our data also supports further research of gathering and analyzing additional similar data to test the relationships or develop and test new pre-test measures.

Implications and Caveats for Managers

The two types of managerial implications that can be drawn from this study are (1) the implications for advertising practice that come from what we explained about the sales effectiveness of television advertising for packaged goods and (2) the managerial implications from what is still unexplained. It has been our experience that managers are more likely to follow the first set of implications, than the second. The second set of implications involves managing uncertainty, an activity managers do not typically enjoy. We now summarize the implications for managers from what we did learn:

1. *T.V. advertising weight alone is not enough.*
   There is no simple correspondence between increased T.V. advertising weight and increased sales, regardless of whether the increased spending is compared to competition or not. This conclusion is consistent with the summary of 19 matched market controlled Campbell's Soup advertising experiments reported by Eastlack and Rao (1989) and those split cable experiments summarized by Blair (1987). Eastlack and Rao concluded that without change in creative executions, changing the spending level had little or no impact on sales for established brands.
2. *The status quo is not enough.*
   T.V. advertising is more likely to work when there are changes in:
   - •Brand/copy strategy
     An increase in sales effect is more likely when (1) the brand's objective is to increase penetration, (2) copy strategy is changed, or (3) copy is intended to change attitudes.
   - •Media strategy
     Standard flighted media plans are less likely to help generate sales increases.
   - •Category dynamics
     Brands in categories in a growth mode or in categories with more purchase opportunities are more likely to be able to improve sales through increased T.V. advertising weight than other brands.
3. *Higher levels of trade display appear to correspond with a reduction in the ability of T.V. advertising to positively affect sales.*
4. *It is unlikely that there is a strong relationship between standard measures of T.V. commercial recall and persuasion for established brands and the sales impact of the copy.*
5. *New brands or line extensions tend to be more responsive to alternative T.V. advertising plans than established products.*
6. *These data support the importance of introductory weight and prime time for new products.*
   Higher boosts in prime time T.V. advertising are correlated with larger increases in sales for new products. A heavy-up plan which is proportionately less front loaded than the base plan results in a reduced likelihood of increased sales for new products.
7. *Concentration of higher T.V. advertising weight is related to increases in brand sales.*
   Weight added to either the front or back end of a media plan is more likely to result in higher sales than increased weight distributed throughout the plan.

In terms of what we "did not learn,"—depending on assumptions one wants to make about shared variance in our data— the data explain less than half of the variance of sales changes associated with T.V. advertising weight changes. If a manager is aware of this uncertainty, he or she can help manage it.

One obvious conclusion is that more real world tests or experiments need to be undertaken. If 61% of the current advertising is not responsive to weight changes, then it would make sense as a manager to run tests to find out if indeed the current campaign is still working on an incremental sales basis. Split cable tests or matched market tests using scanner data can be used to test either alternative weights to the national plan or alternative copies to the current execution. As a manager, it makes sense always to be estimating the incremental effects of the current campaign in the real world.

A particular procedure using experiments as "lead markets" for weight reduction seems appropriate but, to our knowledge, has not yet been implemented in practice. The basic idea is to lower or eliminate T.V. advertising in experimental lead markets. If there is no sales reduction in the lower weight experimental cells after a period of six to twelve months, then lower the weight to the experimental level nationally. Simultaneously, keep the test cells comparing the old current level to the reduced level. Thus, the experimental cells are "lead markets," because they are six to twelve months ahead of the nation in measuring the effects of the reduced levels of T.V. advertising. If the lead markets begin to show a decline in the experimental cells with reduced weight, then the national campaign can be increased back to the old "current" level with a six to twelve month cushion for risk reduction. Simultaneously with the "lead market" approach, the manager should be developing and testing in-market new copy or executions that might be an improvement (on an incremental sales basis) over the current executions.

If the increased productivity of in-market testing is adopted by many packaged goods advertisers, then the role of their sales force will change with respect to its interaction with their advertising programs. Currently, one role of a large advertising budget is to use it as a lever by which the sales force can convince retailers that there will be a lot of consumer pull demand for their products and that the retailers should allocate more shelf space to them. As retailers get more sophisticated, they will begin to ask for evidence that the advertising is actually delivering incremental pull demand in experimental situations. The sales force that can discuss experimental evidence, including covariate adjustment and other technicalities, will have to be more analytical than the typical sales force of today.

Above all, this study has evaluated many of the time honored rules about T.V. advertising and found that many are not to be relied on with the weight currently applied to them.

The experiments clearly show that there are many aspects of advertising sales effectiveness that seem to be unique to a particular brand, competitive situation, copy strategy and execution, and media strategy and tactics. In particular, the opportunity costs of staying with the status quo present a challenging opportunity for increasing advertising productivity.

APPENDIX

The BehaviorScan® system, a service of IRI, is a real-life laboratory in which manufacturers test marketing variables within a controlled environment for both new and established brands. The current BehaviorScan® markets are Pittsfield, MA; Marion, IN; Eau Claire, WI; Midland, TX; Grand Junction, CO; and Cedar Rapids, IA. Past BehaviorScan® markets utilized for tests contributing to this study included:Williamsport, PA; Rome, GA; Salem, OR; and Visalia, CA. The markets are large enough to test in, but small enough to allow precise control of testing variables and product distribution, and virtually complete coverage of grocery store product purchasing. BehaviorScan® testing occurs at the consumer level. Consumer purchasing information is electronically collected from each representative household that has been recruited and maintained in each BehaviorScan® market. Panel members shop with an ID card which is presented at checkouts in scanner-equipped

grocery and drug stores, allowing IRI to electronically track over time each household's purchasing, item by item.

For tests of alternative T.V. advertising plans, the BEHAVIORSCAN® household panels are split into two or more subgroups that are balanced on past purchasing, demographics, and stores shopped over a one-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 pre-existing differences between groups or to an interaction between the test variable and pre-existing differences. After a suitable period of time (usually a year), the test is ended and the results are analyzed using analysis of covariance to remove any consistent influence of uncontrolled factors (e.g., competitive and test brand promotions), so that the impact of the test treatment can be seen more clearly.

Using split cable technology, commercials can be substituted at the individual household level, allowing one subgroup to view a test commercial, while the other views a control ad, or enabling one subgroup to receive heavier advertising weight than the other. In all of the BEHAVIORSCAN® markets, cable is the most viable way of receiving an acceptable television signal. In each market, IRI maintains permanent warehouse facilities and has an in-market staff to control distribution, price, and promotions. Competitive activity in all categories is monitored and a complete record of pricing, displays, and features permits the covariate adjustment of differences that occur across stores.

The data are treated as a post-test/pre-test with control group design. Because the weekly total store scanner data are available along with merchandising data (e.g., newspaper features, displays, price reductions) for every store in which the panelists typically shop, a different analysis method than McGuire's (1977), which was at the individual panel member level, has been shown to work quite well. 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, and price reductions, and regular price reductions, which vary across stores during the same week. In order to accomodate for this source of variation, a co-variate is used—simply the market share for the brand in question for the total store for that week.[2] This variable accounts for store changes at the store week level that are associated with the treatment effect. For example, if a store has a promotion, and brand sales double, then the covariate would isolate that effect from any treatment effect on the panelists who happened to shop that week. Because the panelists are a small fraction of the store's volume, this variable is an elegant, powerful covariate that screens all variables that might affect store week sales. This covariate alone typically explains 20–40% of the variance in store-week-treatment sales over the one-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 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 straight forward demographics and past purchase variables. In all results described, the sales volume and share results are covariate adjusted using the above procedure.

[2]Gerald Eskin was the originator of this powerful variable.

REFERENCES

Aaker, David and J. M. Carman (1982), "Are You Overadvertising?" *Journal of Advertising Research*, 22 (4), 57–70.

Aaker, David and R. G. Myers (1988), *Advertising Management*. New York: Prentice Hall.

Abraham, Magid A. and Leonard M. Lodish (1990), "Getting the Most Out of Advertising and Promotion," *Harvard Business Review*, 3 (May/June), 50–60.

Assmus, Gert, J. W. Farley, and D. R. Lehman (1984), "How Advertising Affects Sales: Meta-Analysis of Econometric Results," *Journal of Marketing Research*, 21 (February), 65–74.

Blair, Margaret H. (1987) "An Empirical Investigation of Advertising Wearin and Wearout," *Journal of Advertising Research*, (December/January), 45–50.

——— and K.E. Rosenberg, (1994), "Covergent Findings Increase our Understanding of How Advertising Works," *Journal of Advertising Research*, 34 (May/June), 35–45.

Deighton, John, Caroline M. Henderson, and Scott A. Neslin (1994), "The Effects of Advertising on Brand Switching and Repeat Purchasing," *Journal of Marketing Research*, 31 (February), 28–43.

Dutka, Solomon (1984), "Combining Tests of Significance in Field Marketing Research Experiments," *Journal of Marketing Research*, 21 (February), 118–19.

Eastlack, J. O. and A. G. Rao (1989), "Advertising Experiments at the Campbell Soup Company," *Marketing Science*, 8 (1), 57–71.

Farley, John V. and Donald R. Lehman (1986), *Meta-Analysis in Marketing: Generalizations of Response Models*. Lexington, MA: Lexington Books.

Haley, Russell I. and Allan L. Baldinger (1991), "The A.R.F. Copy Research Validity Project," *Journal of Advertising Research*, (April/May), 11–32.

Jones, John P. (1989), *Does It Pay to Advertise?* Lexington, MA: Lexington Books, 298–99.

Mahajan, V. and E. Muller (1986), "Advertising Pulsing Policies for Generating Awareness for New Products," *Marketing Science*, 5 (2), 89–106.

McGuire, Timothy W. (1977), "Measuring and Testing Relative Advertising Effectiveness with Split Cable TV Panel Data," *Journal of the American Statistical Association*, 72 (December), 736–745.

Rhodes, R. (1977), "What ADTEL has Learned," paper presented at the AMA Advertising Research Conference, New York, (March 22).

Sasieni, M. W. (1971), "Optimal Advertising Expenditures," *Management Science*, 18, 64–72.

——— (1989), "Optimal Advertising Strategies," *Marketing Science*, 8 (Fall), 358–70.

Simon, H. (1982), "ADPULS: An Advertising Model with Wearout and Pulsation," *Journal of Marketing Research*, 19 (August), 352–63.

——— and J. Arndt (1980), "The Shape of the Advertising Response Function," *Journal of Advertising Research*, 20 (August), 11–28.

Spence, M. A. (1980), "Notes of Advertising Economies of Scale and Entry Barriers, *Quarterly Journal of Economics*, 95 (November), 493–507.

Strong, E. C. (1977), "The Spacing and Timing of Advertising," *Journal of Advertising Research*, 17 (December), 25–31.