

Experimentation in the 21st Century: The Importance of External Validity

Russell S. Winer

University of California at Berkeley

Much of the consumer behavior literature is devoted to what has been referred to as theory applications (TA) research in which the main focus is on laboratory experiments with student subjects and high internal validity. In this article, the author argues that external validity concerns should be given more attention, particularly in TA research. Three recommendations are made for implementing these concerns: (1) consumer behavior articles should be required to have a section indicating how increased levels of external validity can be obtained with other studies, (2) "joint ventures" between consumer behavior and marketing science researchers can be profitable and should be encouraged, and (3) analyses of electronic scanner panel data or other secondary data can be used to generate higher levels of external validity. Three examples are given from the marketing literature of how findings from experiments and scanner data can be combined to advance a stream of research.

One of the most nagging issues facing researchers interested in consumer behavior is that of external validity. External validity, of course, deals with the issue of generalizability of the results found to other populations, settings, and so forth (Campbell and Stanley 1963). One manifestation of this concern about external validity is that, periodically, an article appears in a leading marketing journal excoriating consumer behavior researchers for using student subjects. Since students are not "real" people, how can the lab results be generalized to the greater

population? Such articles are invariably followed by responses that emphasize the fact that some researchers are interested in "theory," not generalization, and for their purposes, student subjects are just fine. For these researchers, what is critical is internal not external validity. As long as we have confidence that the results are truly due to what has been manipulated, what difference does it make if the subjects are "real" men, women, or, as one marketing academic has put it, "little green men from Mars"?

Internal validity is, of course, a necessary condition for any experimental study. External validity is not of much concern in experimental work if the researcher cannot adequately show that the results found from an experiment are truly due to the manipulation(s). However, as Wells (1993) has put it, "We have inherited a myth that says that internal validity is adequate and external validity, if any, is up to someone else" (p. 492). This attitude toward external validity is manifested in articles published in the *Journal of Consumer Research*, the *Journal of Consumer Psychology*, the *Journal of Marketing Research*, this journal, and other major marketing journals publishing empirical work. Rarely does an author of an experimental study either worry about how to establish external validity for the results or actually perform additional studies that go a long way toward establishing some degree of external validity.

There are several theses in this article. First, I argue that as we move into the twenty-first century, the current state of business schools in which marketing academics reside will require more and more that our research is not only of high quality but relevant. By relevant, I do not mean that research has to directly inform practicing managers. However, I believe that it is incumbent on us to be concerned about the generalizability of research results beyond the lab into other contexts. This gives practitioners who are interested in applying our work (and there is much more

even today that can be applied than is) to their problems confidence that empirical results apply to more than 18- to 22-year-olds at "large midwestern universities."

Second, I propose that experimental articles focusing on internal validity in controlled, laboratory environments have a mandatory section at the end of each article indicating what kind of studies are necessary to establish external validity. This will put pressure on consumer behavior researchers to think about external validity as an integral part of their work and to not leave it totally "up to someone else." As part of this proposal, I strongly encourage consumer behavior researchers and marketing scientists, the main group of marketing academics interested in scanner panel data, to work more together rather than go their separate ways, as is currently the situation.

Third, I argue that there are readily available sources of information that can be used to extend the generalizability of many consumer behavior studies—scanner panel data. These data have been around for about 40 years and are a very rich source of information that has been rarely exploited for this purpose. Scanner data represent observations of purchasing behavior of individuals in a real environment. While the environment itself is not the critical factor (we are interested in external not ecological validity), scanner data studies that support results found in the lab provide strong supporting evidence of external validity.

THE EXTERNAL VALIDITY CONTROVERSY

Researchers have developed three perspectives on external validity (Lynch 1982). One perspective is *statistical* generalizability, in which the main issue is whether the results from a study using a particular sampling approach can be generalized to the larger population of interest. *Robustness* is whether a relationship found in an experiment could be replicated with different subjects, research settings, and time intervals. *Realism* (or ecological validity) is whether the research study (tasks, stimuli, settings) was realistic and, therefore, the results likely to be generalizable to a more natural environment. For the purposes of this article, I am particularly interested in the latter two perspectives, although I acknowledge the fact that any one study may not address either or both of them. Rather than worrying about every single study satisfying both internal and external validity criteria, I am more interested in a stream of research that ultimately incorporates both perspectives leading to the more general notion of external validity.

A number of articles appearing in the consumer behavior literature have shaped the discussion about external validity in experimental contexts. I review these briefly below.

Ferber (1977)

In this editorial, Ferber argued that convenience samples, mainly students, should be avoided for two reasons. First and more important for this article, students in the sample may not actually be consumers for the products or services being tested. For example, I have read articles that use televisions and other expensive durable goods as stimuli with student subjects. My guess is that 18-year-old students are infrequently in the market for televisions. The second point is that convenience samples are obviously not probability samples randomly drawn. This criticism applies to a wide variety of marketing studies, not just laboratory studies with student subjects.

Calder, Phillips, and Tybout (1981)

This article really kicked off the controversy in the 1980s. Calder et al. (1981) drew a distinction between two kinds of research: *effects application* (EA) research, in which the researcher is interested in generalizing the results to other settings and populations beyond the current research setting, and *theory application* (TA), in which the theory itself is expected to generalize and not the particular effects or empirical results. An example of EA research is the relationship between price and perceived quality. Marketing managers are interested in whether such a relationship exists and its boundary conditions (e.g., product categories, price ranges). To have confidence that such a relationship exists, empirical results showing that a positive relationship between price and perceived quality must be generalizable to the real world, that is, beyond the experimental research setting. An example of TA is Petty and Cacioppo's (1981) elaboration likelihood model (ELM). In this case, it is not important that the results of a particular study are replicable but only that other studies replicate the theoretical underpinnings of the model, that is, that the forces of persuasion are different when some subjects elaborate on a stimulus versus others that do not.

The authors advocate that there are implications of this distinction on the selection of respondents, operationalizing the independent and dependent variables, the research setting, and the research design. In particular, EA studies need respondents that represent their real-world counterparts, while TA can use any respondent population, preferably as maximally homogeneous as possible for a strong test of the theory. Variables used in EA studies need to correspond as closely as possible to the real world, while TA variables must correspond to the needs of the theory. The research setting used for EA research also needs to correspond as much as possible to the contexts in which generalizability is desired, whereas for TA research, the setting can be artificial as the goal is to create an environment free of sources of extraneous sources of variation that could negatively affect internal validity. Finally, true

experimental designs are preferred for TA research, while EA research can use any design that is again appropriate for the real-world context including "natural" experiments.

This distinction between EA and TA research is important.¹ Calder et al. (1981) would feel that it is important for price-perceived quality research be done in a way that generalizes to the real-world contexts of interest, that is, that there are some product categories in, say, a supermarket, where we would expect to see that kind of relationship. However, they would feel that it is not important to show that consumers actually use the ELM model when making purchase decisions—only that the ELM model is not rejected as a theoretical explanation for observed laboratory behavior. In general, EA studies show some concern for external validity, while TA, none at all.

Lynch (1982)

Lynch disagreed with the Calder et al. (1981) position on external validity.² One of his major points was that if any research findings lack external validity, the theory lacks construct validity. In particular, he noted that a researcher must distinguish effects of the independent, manipulated variables from interactions of the independent variables with background factors that are supposed to be irrelevant. These background factors include subject and setting factors. Thus, if a theory has been tested using only student subjects from one particular geographic area, the researcher does not know if the results would be affected by older subjects or even students from another part of the country or world. Such interactions are threats to external validity. In addition, authors should consider the boundary conditions for their findings, which, besides including omitted background factors, could involve stretching the limits of the levels of manipulated variables.

Lynch (1982) developed three different approaches for designing experimental studies that maximize external validity. The first approach is to allow the different background factors to vary in an experimental design sense and then control for these interactions when analyzing the data (for more detail, see Lynch 1982). The problem with such an approach, of course, is that there might be a very large number of such background factors resulting in an unwieldy and very expensive experiment. A second approach is to be more selective and develop an a priori notion of which background factors are most likely to interact with the treatment variables. The third approach is to deliberately attempt to maximize heterogeneity in terms of respondents and study settings that fall within the domain of the theory through replication. Thus, the researcher might run one experiment using tightly controlled conditions in a laboratory, student subjects, and

running shoes and another using adults, automobiles, and a mall-intercept setting. Assuming the results replicate, confidence in the findings is significantly higher.

Lynch (1982) also makes a comment about how research in an area should progress:

For a given experiment to contribute to progress, someone—whether the original researcher or others in the field—must attempt to replicate it conceptually at some later point in time. Given the low incidence of conceptual replication in our discipline, we undoubtedly would benefit if published research always included some small attempt to test the generality of the findings reported. (P. 237)

Later Contributions

The controversy about the need for external validity in consumer research continued in the pages of the *Journal of Consumer Research*. Calder, Phillips, and Tybout (1982) developed a rejoinder to Lynch's (1982) article indicating that it was impossible to run enough experiments or to sufficiently maximize heterogeneity in experimental designs to ensure that TA findings had external validity. In addition, they felt that research progress did not require any attempt at external validity in any one study. McGrath and Brinberg (1983) attempted to find common ground between the parties while putting forth their own conceptions of how research progress can be made. They develop a three-stage research process in which the final stage is where the researcher or colleagues look for robustness of the findings by systematically varying one or more of the domains of the study (much like Lynch's suggestion to maximize heterogeneity). They also make an important point: it is impossible to increase the external validity of a given study within that study. The external validity of a study can only be assessed in terms of results of another study or series of studies.

Finally, Wells (1993) criticizes the progress that has been made by consumer behavior researchers toward achieving the goals set for the discipline in the early 1970s by pioneers such as Jagdish Sheth, Ronald Frank, and others. He noted, for example, that Sheth called for research to be done in "naturalistic and realistic" settings as far back as 1972. Wells suggested five guidelines for breathing new life into the field. One of these, "forsake methodology," is particularly appropriate for this article, especially two "myths." The first is that "students represent consumers." This is, of course, old ground. The second is that "the laboratory represents the environment." Lab studies do not represent the environment due to the control available to the researcher that is unavailable in the real world, the fact that experimental choices have no short- or long-term consequences for the subjects, the possible existence of demand

effects, and the fact that experiments, unlike the real world, have a sudden beginning and a sudden end.

RECOMMENDATIONS

I feel that Lynch (1982) is correct in his urging of consumer behavior researchers to be more concerned about external validity, however you wish to define it. In addition, I agree in principle with his recommendation that even TA studies should seek some generalization, although not necessarily in the same study. I therefore make three recommendations for furthering consumer behavior research in the twenty-first century.³

My first recommendation is that every consumer behavior article, both EA and TA, have a section at the end discussing external validity concerns and suspected boundary conditions that, of course, limit external validity. Obviously, EA articles have a greater degree of external validity due to their research objectives and design. In this case, the authors should discuss the limitations of the external validity generated from the study. However, the key point is that researchers more interested in TA research should not be absolved from external validity concerns. These authors should either (1) combine their TA-focused experiments with some other experimental or quasi-experimental design in the same article or (2) develop a detailed description of the kind of study or studies necessary to develop a greater confidence that it is not only the theory that can be replicated and generalized but the empirical results as well.

Why is this focus on external validity more important in the future than today? Given the amount of attention this topic has received, some would argue, of course, that it has and should always been important. Researchers primarily interested in TA research without concerns for external validity are, of course, some of the best academics in marketing and highly valued colleagues to all of us. However, there is a big distinction between being a social or cognitive psychologist in a psychology department versus being a member of a marketing department or group in a business school.

As we all know, business schools have been under increasing pressure to be "relevant" to the business world. This need to be relevant has resulted in a number of changes in business school faculty hiring practices, for example. It is no longer enough to be only a promising scholar to get a job at a top school. The most sought-after candidates in the late 1990s also must have the potential to be excellent teachers so that deans can mollify impatient MBA students. This is sometimes referred to as the *Business Week* effect, resulting from the magazine's biannual survey of alumni and recruiters resulting in rankings of the top U.S. schools. In addition, business school communications offices are developing newsletters and other

materials for distribution to their constituents "translating" faculty research into terms they can understand and, perhaps, even use.

While I deplore many of the *Business Week* effects (for example, new business school deans stating their goal to be a "top 10" school), I believe that the pressure on faculty to ultimately develop more relevant research is well placed. Marketing academics should be as well trained in their basic discipline (usually psychology or economics) as possible, and this training should be demonstrated in their research. However, most marketing academics chose marketing doctoral programs and to concomitantly take faculty positions in business schools rather than social science departments. This choice not only implies that we have to teach students who are more interested in the real world than the laboratory world but also that we have to think and should be interested in thinking about our research in the same way.

Note that this does not mean that I think marketing academics should become consultants. In addition, I am not calling for a ban on TA research. Our outstanding consumer behavior scholars primarily interested in TA research should be encouraged to continue to do it. My point is that given that we have chosen to be business school faculty and the increased pressure on business schools to produce students and research that informs practitioners, our research should at least point the way toward more generalization of empirical findings than is currently the case.

Even with some thought given to the problem, it is, of course, cheap and easy to pursue my second alternative, that is, simply to recommend to others how external validity might be obtained. This is, as Wells (1993) put it, really only making it someone else's problem. However, I also recognize the limits that training puts on a researcher's skills, that is, one's "comparative advantage." Scholars trained to do TA research are not necessarily familiar with other data or research methods that could be applied to the same research problem and provide a considerable amount of external validity to the work. In addition, scholars from every discipline normally develop a research "routine" in which they tend not to stray too far from what has made them successful. Another way to say this is that there is a considerable amount of inertia in research programs pursued by academics.

How can we break out of these routines? My second recommendation is that more "joint ventures" should be sought between consumer behavior researchers and people with other disciplinary approaches in marketing. Excellent candidates for the latter are marketing scientists. There are a large number of marketing scientists who are interested in consumer behavior but who attack problems from the perspective of another tradition. I consider myself to be of this group. Rather than running tightly controlled experiments, marketing scientists are more likely to use

scanner panel or other secondary data to test consumer behavior hypotheses. Often, the tests involve specifying alternative models of consumer decision making, estimating the models, and then choosing the model with the best fit or out-of-sample predictions as being most consistent with the specified behavior.⁴ Alternatively, estimated values and statistical significance of the parameters of the models are interpreted as providing evidence of the underlying consumer behavior.

Secondary data sources such as scanner panel data are particularly appropriate for assessing external validity for a wide variety of consumer behaviors. Scanner data and its predecessor, diary panel data, have been analyzed by marketing academics for 40 years, beginning with Kuehn's studies in the late 1950s of the brand choice process (Kuehn 1962). Scanner data present the researcher with actual consumers making purchases in their real environment, the supermarket. The data are collected in an unobtrusive way simply by scanning in a bar-coded panel membership card that identifies the household and subsequently the purchases made on that shopping trip through the bar codes on the products. Researchers obtain measures of brand choice, quantity, price paid, promotions used, time of day, day of week, store choice, and several in-store "causal" variables such as whether a particular brand was being featured in the store. Some scanner panel data sets also provide measures of television advertising exposure ("single source" scanner data) and other measures such as whether a brand was featured in a newspaper during a particular week.

Scanner data are not perfect. We do not know which person in the household is making the purchases (except, of course, for single-person households). This is important: for food items, multiple brand or flavor purchases can represent different household preferences that are unknown. We also do not have any consumption data. In addition, while the samples are much more representative than they were using the old diary technology and the data are collected easily with no effort on the part of the panel members, there are always questions about the kind of people who agree to be on these panels as well as the "mortality" issue of panel dropouts. Finally, and important for consumer behavior researchers, there are no process measures (e.g., attitudes) taken, as only purchasing behavior is measured.

Despite these problems, scanner panel data represent real people making real decisions in a real environment. These three characteristics uniquely distinguish panel data work from laboratory experiments. It is obvious what the trade-offs are: internal validity (experiments) for external validity (scanner data). However, what scanner panel data offer is more than a realistic setting. If laboratory results hold in an analysis of one or more scanner panel data sets, this gives confidence that the lab results are not likely to

change by varying what Lynch (1982) referred to as background factors. That is, what the realistic setting provides is not just the supermarket but the fact that most of the background factors that a researcher cannot hold constant are at work in the real world. The background factors that cannot be controlled in the lab naturally vary in a supermarket. Different people in the household shop (students, parents, retired people, etc.), products are sold at different shelf heights, babies may or may not be screaming, brands may or may not be on sale, and so forth. If results from the lab can hold in this kind of "dirty" environment, what we have is a strong form of external validity.

Thus, my "ideal" form of the research process is a lab experiment in conjunction with a "natural" experiment like scanner panel data, my third recommendation. Note that the process does not have to work in the direction experiment → scanner data. It is also possible for results found in scanner data studies to be given internal validity using lab experiments. I do not expect the same person to do both kinds of work; however, partnerships between scholars trained from different perspectives can and have brought complementary insights and skills to bear on a research problem that has resulted in greater generalizability to the results than would have resulted from either alone.

ILLUSTRATIONS

I offer three examples of such partnerships that show how laboratory experiments and scanner panel studies can be used complementary for the purposes of generating external validity.

Simonson (1990)/Simonson and Winer (1992)

The research question posed by Simonson (1990) was, How does purchase quantity and timing affect variety? In particular, one of the main hypotheses was the following:

Consumers who simultaneously choose multiple items in a category for sequential consumption are more likely to choose different items than consumers who sequentially make the same number of choices.

There were several motivations behind this hypothesis. The desire for a "varied experience" has been widely found to be an important driver of purchase and consumption. This is likely to be true when consumers are uncertain about future preferences, particularly when multiple purchases are made simultaneously. Choosing variety also reduces risk and is an efficient strategy when someone is having difficulty making a decision.

Several studies were done to test this hypothesis. The studies all used undergraduates in either business or psychology courses. The settings were classrooms or similar environments. There was some variation in the reality of the setting as one of the studies involved real choices (the subjects did not have to spend their own money) over a period of 3 weeks versus simulated choices based on the subjects' imaginations. In addition, there was some variety over product categories across the studies. The experiments all confirmed the above hypothesis in that, for example, subjects who simultaneously chose three snacks for three future consumption occasions were more likely to select a variety of items than were those who chose three snacks sequentially, one on each consumption occasion.

As a result, we have an interesting finding with high internal validity but little external validity. Student subjects in laboratory environments did indeed exhibit behavior supporting increased preference for variety when making simultaneous versus sequential choices. Although some managerial implications of the research were discussed (as usual), no attempt was made to generalize the findings to other settings or populations.

To extend these laboratory results, Simonson and Winer (1992) analyzed a scanner panel data set from the yogurt product category. Clearly, the kind of control available in the lab is not available with such secondary data. With scanner data, the researcher has to do the best job possible replicating the lab environment in an *ex post* fashion. A key issue was how to define variety within a product category rather than between categories. If the key dependent variable cannot be replicated, then the scanner panel study may be producing interesting results, but they are not producing results useful for the purposes of external validity.

In this study, variety was defined based on the frequency with which a particular flavor was purchased over the entire purchasing history available. Each household purchase of a unit of yogurt was assigned a variety index based on that flavor's share of all purchases. As a result, smaller variety indexes were associated with flavors purchased less often. The "amended" hypothesis was the following:

As the number of items purchased in a category on a given occasion increases, consumers are more likely to choose flavors that they do not usually buy when making fewer purchases.

In other words, the study used a within-household design in which the hypothesis would be supported if more purchases on an occasion led to flavors with lower variety indexes being purchased.

Using households with more than 10 purchases of the category, regressions were run for each household with

each purchase of an item counted as one observation. The model was as follows:

$$\text{Variety index} = f(\text{number of items purchased on that occasion, price, promotion}).$$

Price and promotion were included to control for their effects. The hypothesis would be supported if the coefficients across households on the number of items purchased were significantly negative. Of the 1,694 coefficients (one for each household across purchase occasions), 63 percent were significantly negative and 37 percent were either insignificant or positive. To control for intrahousehold taste differences, the results were replicated on single-person households. Thus, the scanner panel data results supported the lab studies: greater purchase quantity on an occasion leads to greater variety chosen.

Kahn and Raju (1991)

In this article, the authors draw a distinction between variety-seeking consumers in which the previous purchase decreases the probability of the same brand being purchased on the next occasion and reinforcement consumers in which the previous purchase increases that same probability. The issues examined are how the frequency of price discounts affects the choice behavior of these two groups and how the effects of discounts are mediated by whether the brand in question is a major (large share, base probability of purchasing $>.5$) or minor ($p < .5$) brand.

Based on a mathematical formulation and assuming that the market consists of two brands with only one brand promotion at a point in time, the authors develop two propositions:

1. For all promotional frequencies (including no promotion), the probability of choosing a major brand is higher for reinforcement consumers than for variety-seeking consumers. However, an increase in the frequency of discounts has a greater impact on purchase probabilities for variety-seeking consumers.
2. For minor brands, when promotions are infrequent (including no promotion), the long-run choice probability for a minor brand is higher for the variety-seeking segment than for the reinforcement segment. However, when promotions are frequent, the long-run choice probability of choosing the minor brand is the reverse, that is, it is lower for the variety-seeking segment than for the reinforcement segment. In addition, an increase in the frequency of discounts has a greater impact on purchase probabilities for reinforcement consumers.

An implication of the second proposition is that, for minor brands, there is an interaction effect between the frequency of promotion and the type of brand. Taking the first and

second proposition together implies a three-way interaction between (a) purchasing behavior (reinforcement or variety seeking), (b) brand (major or minor), and (c) whether the brand of interest is promoted.

These implications of the model were tested in a laboratory experiment run using 25 undergraduate business students. Using a within-subject design, subjects were asked to make a series of choices between two hypothetical brands of furniture polish. Each brand was described by its attributes, brand name, and price. Subjects had to make 6 sets of 20 brand choices representing two brand choice conditions (variety seeking and reinforcement) by three promotion conditions (no promotion, promotion only on one brand designated as the major brand, promotion only on the minor brand). To simulate the brand choice conditions, subjects were told either that repeated polishings of a desk by the same brand were necessary for maximum benefit or that switching was necessary to avoid wax buildup.

The findings generally supported the propositions. In the no-promotion condition, the market share of the major brand was larger under the reinforcement condition than under the variety-seeking condition; the results were the opposite for the minor brand. For the minor brand, increasing the frequency of promotion (from no discount to a discount) had a relatively larger effect in the reinforcement condition than in the variety-seeking condition; the opposite effect for the major brand was insignificant. Finally, the hypothesized three-way interaction was significant.

These results were extended using scanner panel data from the cracker product category. Two brands were selected from the saltine subcategory, one with a significantly larger share than the other. In the flavored cracker subcategory, five brands were analyzed excluding private labels. In both subcategories, panelists were assigned to either the variety-seeking or reinforcement behavior segments based on their purchase patterns in periods of low promotion. Thus, like the previous illustration, the lack of control is manifested by the inability of the researchers to randomly assign subjects to the different treatment groups; that is, the experiment is "natural." Again, while the effects for the minor brands were somewhat weaker for the flavor cracker subcategory, the propositions developed from the authors' theory were supported by the scanner panel data analysis. These results clearly provide external validity for those found in the laboratory experiment.

Leclerc and Little (1997)

In this study, the authors examine the role of advertising copy in free-standing insert (FSI) promotions. In particular, the authors are interested in how the creativity of the advertising copy accompanying the clip-out portion of the promotion affects attitudes toward the promoted brands and how these effects might vary over two segments

targeted by such promotions: customers loyal to competitors' brands and brand switchers.

The theoretical arguments behind the authors' propositions are based on work that has found that a person's motivation to process arguments moderates persuasion. When consumers examine an FSI, their motivation to process the information will depend on their brand loyalty or commitment and the level of involvement in the product category. Because of their preexisting attachment to a brand, loyal customers have little reason to process the information in an FSI. On the other hand, brand switchers have not made a decision about which brand to purchase and are therefore more likely to be motivated to examine the information in an FSI, particularly in high-involvement product categories. Neither group, loyals or switchers, is likely to be motivated to process FSI information in low-involvement categories.

The authors first conducted an experiment with three kinds of FSIs, each with a basic product display and a headline: (a) the product display and headline only, (b) the first ad plus brand information, and (c) the first ad plus an attractive picture. The category was a high-involvement beverage (cranberry juice). Based on the theory, brand-loyal consumers should not be affected by advertisement (b), as they have low motivation to process the information. However, because advertisement (c) did not involve any processing but did contain an attractive peripheral cue, loyal customers should be influenced positively toward the advertised brand. Brand switchers should find advertisement (b) more appealing, as they are more motivated to process information. However, they should not find (c) appealing, as they are motivated to process information and do not find any in that advertisement associated with the FSI. Finally, advertisement (a) should not have any impact on brand attitudes.

More formally, the first experiment tested the following two hypotheses:

Hypothesis 1A: Brand loyalty interacts with executional cues: for customers with high loyalty to a competitive brand, an advertisement featuring an attractive picture and no target brand information will generate a more positive attitude (and higher propensity to clip) than an advertisement providing brand information. For customers with low loyalty (switchers), an advertisement featuring brand information will generate a more positive attitude (and a higher propensity to clip) than an advertisement featuring an attractive picture and no brand information.

Hypothesis 1B: Compared to an advertisement featuring product display only, advertisements featuring executional cues will generate a more positive attitude toward the brand (and a higher propensity to clip).

A laboratory experiment was conducted to test these two hypotheses. The subjects, who averaged 37 years of

age, were staff members of a university. The experimental brand was real, and the subjects were asked to evaluate FSIs similar to those actually found in Sunday newspapers. Subjects were screened for perceptions of the category as high involvement and for being nonusers of the focal brand (since the focus of the study is on brand switchers and consumers loyal to a competing brand). The subjects were assigned randomly to one of three experimental conditions: (1) information-oriented advertisements, (2) a pleasant picture with no brand information, and (3) a picture of the package only (no product information and no pleasant picture).

The experimental results supported the hypotheses. When loyalty to the competing brand was high, the advertisement with the attractive picture generated more positive brand attitude than the advertisement with brand information. In addition, as loyalty decreased (switching propensity increased), the information-oriented ad worked increasingly well. The result supporting Hypothesis 2, executional cues are superior to product display only, was in the right direction but not significant. With respect to actual clipping behavior, all the results were in the appropriate direction but statistically insignificant. Overall, there was reasonably strong support for the hypothesized interaction between copy execution cues and loyalty in the one category studied.

The authors attempted to support the general conceptual framework (motivation to process information moderates persuasion) using cross-sectional data measuring FSI behavior in a scanner panel. With these data, loyalty measures are obviously not available, but there were multiple categories and opportunities to measure degree of involvement with those category purchases. Thus, the authors attempted to measure the interaction effects of involvement and executional cues on FSI redemption behavior. The latter was measured by what is termed coupon efficiency, the proportion of coupon redemptions representing incremental sales.

Although the previous studies focused on high-involvement products, the framework also predicts that advertisements with brand information should not affect attitudes for switchers in low-involvement product categories since switchers are not motivated to process information. Thus, a hypothesis (numbered H3 in the article) is the following:

Hypothesis 3: Brand information will have no effect on coupon efficiency for product categories generating low levels of consumer involvement. As involvement increases, the effect of brand information on efficiency will increase.

The authors studied 387 coupons evaluated by IRI's CouponScan service along with their measured efficiencies for six product categories: cookies, crackers, marga-

rine, fruit juice, bar soap, and breakfast cereals. Each coupon was coded by independent raters for two executional cues: the degree of brand information and the degree of visual elements (7-point scales). Each product category was also evaluated by independent raters for involvement. The model estimated was the following:

$$\text{Efficiency} = f(\text{information, visual elements,} \\ \text{involvement, brand share, coupon face value,} \\ \text{Information} \times \text{Involvement, Visual Elements} \times \text{Involvement}).$$

The two variables, brand share and coupon face value, were included as control variables, and the Visual Elements \times Involvement interaction tested an additional hypothesis not covered in this article.

As predicted, the Information \times Involvement interaction was significant and positive. Also consistent with Hypothesis 3, the main effect of information was insignificant. Thus, the scanner panel analysis adds a strong degree of external validity to the laboratory experiments and, concomitantly, to the behavioral framework developed by the authors.

CONCLUSION

These three illustrations show that there are considerable benefits from collaborations between scholars interested in studying consumer behavior from the psychological perspective and those interested in modeling behavior using marketing science approaches. In all three studies, research hypotheses were first developed using relevant prior literature from psychology and/or marketing, controlled experiments were run, and then the experimental results were generalized using scanner panel data. Interestingly, there was some variation in the experimental work in terms of EA versus TA research. The Leclerc and Little (1997) experimental study was clearly an example of EA, as the design was intended to mimic as much as possible the real-world environment in which consumers clip FSI coupons. A careful reading of the Kahn and Raju (1991) article also indicates that, despite using student subjects and more artificial stimuli, the intent of the experimental work is more in line with Calder et al. (1981)'s notion of EA rather than TA. Simonson's (1990) work, however, is more clearly in the TA camp, as the focus is more on the theoretical explanation for the finding rather than whether the particular effect found would replicate in the real world.

Thus, scanner data were found to be useful for external validity purposes for both EA and TA experimental studies. While the former may have less need than the latter for empirical evidence favoring generalization, both kinds benefit from such work. In addition, it can be seen that it is not critical for both experimental and scanner studies to

appear in the same article. What is interesting is that the EA experimental studies appeared in the same articles as the scanner studies. A conjecture is that the researchers had real-world effects in mind from the beginning that affected the experimental designs.

It is clear that since the early 1980s when the external validity debate appeared in *Journal of Consumer Research*, some movement in the direction of the recommendations of this article have already taken place. Many consumer behavior articles include the following:

- multiple studies in which previous results are replicated and different manipulations, subjects, or procedures are used;
- main-effects experiments followed by studies with interaction effects looking for boundary conditions and reversals;
- measuring subject motivation and involvement with the experimental task and removing subjects who fail these screens;
- increasing the realism of tasks; and
- using covariates to control for individual and situational interactions.

A good example of a study using some of these experimental improvements is Inman, Peter, and Raghubir (1997). In this article, the authors examine the use of purchase restrictions as information used by consumers in evaluating promotions. The authors used three different research methods (grocery sales data, a simulated grocery store experiment, and a survey), different samples (West Coast, Midwest, and Hong Kong), and three different operationalizations of restrictions (purchase quantity limit, purchase precondition, and time limit) to demonstrate that imposing such purchase restrictions consistently increases the choice probability of the restricted brand.

I should emphasize the fact that, although it was not the topic of this article, much can be learned by reversing the order of studies. In other words, following up scanner studies purporting to find some behavioral phenomenon with a lab experiment brings much-needed internal validity to the research stream. This is to be encouraged as well. For example, in Simonson and Winer (1992), the scanner study that followed the lab study was itself followed by an experiment to test some interesting questions raised by the scanner data results concerning the display of products in terms of groupings by flavor or brand. Thus, my argument is not completely one-sided: marketing scientists have an obligation to consider the experimental implications of their work as well.

Thus, I strongly encourage journal editors to consider the policy recommended in this article: behavioral researchers focusing on TA should be asked to give specific recommendations concerning increasing the external validity of that stream of research. Modelers should also be

encouraged to think of how experiments could be designed to support their empirical analyses. In addition, research joint ventures should be strongly supported. One way for this to happen is through doctoral seminars: discussions of TA work could include what kind of secondary data would be useful for the purposes of external validity. Likewise, marketing model seminars can assign students to thinking about experimental work that can support empirical studies using secondary data like scanner data.

Marketing academics should and usually do think about a research area as a stream of research. Recent questions have been raised about exactly what should constitute the "next" study in a research stream from a meta-analysis perspective (Farley, Lehmann, and Mann 1998). In this article, I take the view that the next study, if necessary, should focus on external validity issues. These issues will be central to our roles as marketing academics in business schools as we move into the twenty-first century.

ACKNOWLEDGMENTS

The author appreciates comments made on an earlier draft by Priya Raghubir, Itamar Simonson, and Joydeep Srivastava.

NOTES

1. In practice, it may be difficult to classify a given study as belonging to one group or the other. An effects application study may share many of the characteristics of a theory application study with the difference being focused on the intent to find the experimental results in the real world.
2. I should be clear that Lynch (1982) did not completely disagree with all of the arguments set forth in the Calder, Phillips, and Tybout (1981) article.
3. My major focus is on consumer behavior research using classical scientific methods rather than other approaches such as interpretive methods.
4. See, for example, Stiving and Winer (1997).

REFERENCES

- Calder, Bobby J., Lynn W. Phillips, and Alice M. Tybout. 1981. "Designing Research for Application." *Journal of Consumer Research* 8 (September): 197-207.
- , ———, and ———. 1982. "The Concept of External Validity." *Journal of Consumer Research* 9 (December): 240-244.
- Campbell, Donald T. and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin.
- Farley, John U., Donald R. Lehmann, and Lane H. Mann. 1998. "Designing the Next Study for Maximum Impact." *Journal of Marketing Research* 35 (November): 496-501.
- Ferber, Robert. 1977. "Research by Convenience." *Journal of Consumer Research* 4:57-58.
- Inman, J. Jeffrey, Anil C. Peter, and Priya Raghubir. 1997. "Framing the Deal: The Role of Restrictions in Accentuating Deal Value." *Journal of Consumer Research* 24 (June): 68-79.

- Kahn, Barbara E. and Jagmohan S. Raju. 1991. "Effects of Price Promotions on Variety-Seeking and Reinforcement Behavior." *Marketing Science* 10 (Fall): 316-337.
- Kuehn, Alfred A. 1962. "Consumer Brand Choice—A Learning Process?" *Journal of Advertising Research* 2 (December): 10-17.
- Leclerc, France and John D. C. Little. 1997. "Can Advertising Copy Make FSI Coupons More Effective?" *Journal of Marketing Research* 34 (November): 473-484.
- Lynch, John G., Jr. 1982. "On the External Validity of Experiments in Consumer Research." *Journal of Consumer Research* 9 (December): 225-239.
- McGrath, Joseph E. and David Brinberg. 1983. "External Validity and the Research Process: A Comment on the Calder/Lynch Dialogue." *Journal of Consumer Research* 10 (June): 115-124.
- Petty, Richard E. and John T. Cacioppo. 1981. *Attitudes and Persuasion: Classic and Contemporary Approaches*. Dubuque, IA: William C. Brown.
- Simonson, Itamar. 1990. "The Effect of Purchase Quantity and Timing on Variety Seeking Behavior." *Journal of Marketing Research* 27 (May): 150-162.
- and Russell S. Winer. 1992. "The Influence of Purchase Quantity and Display Format on Consumer Preference for Variety." *Journal of Consumer Research* 19 (June): 133-138.
- Stiving, Mark and Russell S. Winer. 1997. "An Empirical Analysis of Price Endings with Scanner Data." *Journal of Consumer Research* 24 (June): 57-67.
- Wells, William D. 1993. "Discovery-oriented Consumer Research." *Journal of Consumer Research* 19 (March): 489-504.

ABOUT THE AUTHOR

Russell S. Winer is the J. Gary Shansby Professor of Marketing Strategy, the associate dean for academic affairs, and the chair of the marketing group at the Haas School of Business, University of California at Berkeley. He received a B.A. in economics from Union College (New York) and an M.S. and Ph.D. in industrial administration from Carnegie-Mellon University. He has been on the faculties of Columbia and Vanderbilt universities and has been a visiting faculty member at M.I.T., the Helsinki School of Economics, the University of Tokyo, and École Nationale des Ponts et Chaussées. He has written three books, *Marketing Management*, *Analysis for Marketing Planning*, and *Product Management*, and has authored more than 50 papers in marketing on a variety of topics, including consumer choice, marketing research methodology, marketing planning, advertising, and pricing. He is the editor of the *Journal of Marketing Research* and is on the editorial boards of the *Journal of Marketing*, *Journal of Consumer Research*, and the *Journal of Interactive Marketing*. He is the academic director of the Fisher Center for the Strategic Use of Information Technology. He has participated in executive education programs around the world and is an academic trustee of the Marketing Science Institute.