

Theory and External Validity

John G. Lynch, Jr.

Duke University

Winer (1999 [this issue]) proposes that external validity concerns require more attention in theoretical research. The author argues that one cannot “enhance” external validity by choosing one method over another. External validity can only be “assessed” by better understanding how the focal variables in one’s theory interact with moderator variables that are seen as irrelevant early in a research stream. Findings from single real-world settings and specific sets of “real” people are no more likely to generalize than are findings from single laboratory settings with student subjects. Both the laboratory and real world vary in background facets of subject characteristics, setting, context, relevant “history,” and time. It is only when these facets vary and we see how they interact that understanding of external validity is enhanced. For this to happen, the observable “background” factors have to be conceptualized in terms of more general constructs and incorporated as moderators into the researcher’s theory. Enriched theory—not method—confers confidence in our understanding of whether effects will be robust or highly contingent. To map this knowledge to some specific substantive system requires an added step of understanding the mapping from observables in that system onto theoretical constructs. The author proposes “friendly amendments” to Winer’s three proposals to pursue a better understanding of external validity through theory.

Winer (1999 [this issue]) advocates steps that marketing scholars and editors of marketing journals can take to increase the external validity of findings from laboratory experiments testing theories of consumer behavior. First, authors reporting laboratory experiments could include a section in the discussion about “how increased levels of external validity can be secured including boundary

conditions” to the findings reported. Second, “joint ventures” might be encouraged between consumer behavior researchers trained in experiments and marketing scientists trained in econometric analysis of naturally occurring marketplace behavior. Third, some studies might augment the report of one or more laboratory experiments with analyses of scanner panel data.

Winer motivates his recommendations using arguments from past marketing articles on the proper role of external validity in experimental consumer research, including Ferber (1977); Calder, Phillips, and Tybout (1981, 1982); Lynch (1982); McGrath and Brinberg (1983); and Wells (1993). I will point out areas of agreement and curmudgeonly disagreement with Winer’s thesis, drawing from the same articles and from Cook and Campbell (1979), Dipboye and Flanagan (1979), and Lynch (1983).

- I endorse Winer’s suggested discussion section but suggest modifying its slant. External validity can never be “increased,” because external validity is a function of the laws of behavior—that is, whether the manipulated variables combine additively or interactively with a host of background factors. I would therefore propose a friendly amendment to Winer’s proposal, encouraging intelligent, theoretically motivated discussion of the most likely boundary conditions and moderator variables that might change the findings reported. This is a call for speculation about the relationships among latent constructs, and, as such, it is just as important in reports of field studies as of lab experiments. I would add to this a call for discussion about the mapping from levels of the latent constructs to observables. This would help scholars and lay audiences understand how to make educated guesses about the implications of the reported findings to substantive systems of interest to them. Such discussion might be followed by a brief outline of how the boundary conditions might be tested empirically. This would replace the typical contents of “limitations and fu-

ture research” sections. Most such sections present an uninspired rehash of methodological orthodoxy rather than advancing theory. I use Mitra and Lynch (1995) to illustrate a useful external validity discussion in line with Winer’s call.

- I endorse Winer’s proposed joint ventures but suspect that they will continue to be rare. I outline three cases in which it would be worthwhile to make the effort.
- I express skepticism about Winer’s call for augmenting laboratory experiments with scanner panel analyses, except under special circumstances. It requires a very high scholarly investment to master the intricacies of working with scanner panel data. There is little reason to believe that findings from a scanner panel category are highly generalizable across different consumer packaged goods, to say nothing of services, durables, and so forth. Moreover, for most behavioral theories, reliance on scanner panel data is likely to lead to omitted variable bias. The companies that operate household scanner panels jealously guard their panelists and prohibit asking survey questions that might be required to rigorously test most behavioral theories. The work Winer cites is ingenious, but I do not think it can or should become the norm.

My comments reiterate the two major themes of Lynch (1982, 1983). First, neither the theory tester nor the applied marketer can afford to take external validity lightly. For the theorist, the key issue is that unanticipated failures of external validity provide evidence that the theory lacks construct (nomological) validity; the theory is wrong. Interactions of theoretically predicted treatment effects with background factors presumed irrelevant should therefore inspire theoretical progress. For the applied researcher, the incompleteness of one’s theoretical model makes it impossible to know what one should be matching when one attempts to match some test system to some extrapolation system, as in Calder et al.’s (1981) “effects application.”

Second, both rigorous theory tests and appropriate extrapolation from academic research to applied settings of interest depend far more on having an appropriate model of the behavior than on method. Cook and Campbell (1979), say it best. “Indeed, external validity and construct validity are so highly related that it was difficult for us to clarify some of the threats as belonging to one validity type or the other” (p. 82).

ASSESSING OR INCREASING EXTERNAL VALIDITY?

Winer (1999) proposes 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. His second and third recommendations, joint ventures and use of scanner panels, seem to me to be predicated on the view that one increases external validity by conducting studies in the field rather than the lab. My view is that external validity can never be “increased” or “established” because it is a function of the laws of behavior—that is, whether the manipulated variables combine additively or interactively with a host of background factors. Although Winer notes this (p. 351), proposed reforms do not seem to reflect this very important point. Our goal should be to “assess” not to “increase” external validity to better understand where findings do and do not apply.

External validity can only be assessed by better understanding how the focal variables in one’s theory interact with moderator variables that are seen as irrelevant early in a research stream. Findings from single real-world settings and specific sets of “real” people are no more likely to generalize than are findings from single laboratory settings with student subjects. Just as in the laboratory, the real world varies in background facets of subject characteristics, setting, context, relevant “history,” and time. It is only when these facets vary and we see how they interact that understanding of external validity is enhanced. For this to happen, the observable “background” factors have to be conceptualized in terms of more general constructs and incorporated as moderator variables in the researcher’s (now, more complete) theory.

Most researchers, including Winer (1999); Calder, Phillips, and Tybout (1981, 1982); Ferber (1977); and Wells (1993)—and even Cook and Campbell (1979)—associate external validity with methodological procedures: probability sampling, use of “real” consumers rather than students, realistic settings, real behavior, and so forth. I disagree.

One of my two main points in Lynch (1982, 1983) was that these procedures do not have the efficacy in enhancing external validity commonly ascribed to them. The only path to understanding the generality of one’s findings is to have a theory that specifies moderator variables and boundary conditions and that specifies what variables should not moderate the findings reported and to test for the asserted pattern of interactions. If one’s theory is impoverished, no degree of adherence to methodological prescriptions will help “ensure” external validity. One will not have the insight to measure or manipulate the relevant background factors and to elevate them from background variables to theoretical ones. Hutchinson, Kamakura, and Lynch (1998) refer to this as the “ignorance is bliss” problem. If relevant (interacting) background variables vary within cell in one’s experiment, one will not have the insight to block on them so as to avoid aggregation biases (Lynch 1982:227-231). Thus, one will not know whether

one's findings apply at the level of individuals. Hutchinson et al. call this the "some-none-all" problem.

External validity is about theory, not about method. Some methods are better than others, however, at giving one a chance to detect unanticipated interactions with background factors that were not deemed central when the study was designed (Hutchinson et al. 1998; Lynch 1982).

"GENERALIZING TO" OR "GENERALIZING ACROSS"?

Most of my disagreements with Winer (1999) stem from our different definitions of external validity. Winer says, "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)" (p. 349). Cook and Campbell (1979:37, 72-73), though, made an important distinction between generalizing "to" some defined population and generalizing "across" sub-populations of some larger population defined in terms of their levels of various background variables. The first type of external validity they associate with scientific sampling techniques. The second involves freedom of some internally valid effect from interactions with background variables held constant in one's experiment.¹

The point I made in my earlier article was that external validity is really all about "generalizing across" (Lynch 1982:229; see also Cook and Campbell 1979:72-73). We teach our marketing research students that we can guarantee by probability sampling that our research findings will generalize *to* a target population of interest with known levels of confidence. Lynch (1982:226-228) showed why this is pseudo-science. The population elements in statistical sampling theory are not people. They are dependent measures of behavior that happen to be nested in a person—as well as being nested in a level of setting, context, and time. The populations "to" that we would like to generalize all involve measures of future behaviors. The measures of future behaviors have zero probability of being sampled in one's (present) lab or field experiment.

In many behavioral domains of interest to consumer researchers, there is no way of completely enumerating the population of measures from which to sample. Instead, most implementable sampling plans involve sampling on those *independent* variables that we think might affect the estimate—or that we think might serve as proxies for the true causal factors that influence the dependent measure. In such cases, because one's judgment of what must be considered in sampling is fallible, statistical sampling theory can make no rigorous statement about the generalizability of one's results. The reason for this

is that it is impossible to attach a meaningful probability to each observation sampled. (Lynch 1982:228)

Thus, there are no probability samples in basic or applied consumer research; there are, at best, quota samples (cf. Ferber 1977). The trick is to have a theory of what background variables one might use as quota variables. One needs to know what variables might interact with the principal effects demonstrated in a study to be able to project from the test system to some extrapolation population. This is precisely the kind of knowledge relevant to understanding whether one can generalize experimental effects "across" from one stratum to another.

There is one other important reason why generalizing "across" is more important than generalizing "to." Note that even if one somehow managed to sample "randomly" from the true population of interest in an experiment, ignorance of background factors would lead to aggregation biases. The treatment mean effects in the population need not describe any stratum within the population. For example, Lynch (1982:230) described a taste test experiment in which respondents rated their liking for three cola brands on a scale from 1 (*worst*) to 10 (*best*). The researcher is ignorant of moderator variables and so neither measures them nor includes them in the analysis. Unbeknownst to the researcher, the key moderator variable out of hundreds of possibilities is occupation. Students rate Coke, Pepsi, and Shasta as 8, 7, and 4, respectively, and housewives rate the same colas 2, 6, and 7, respectively. If students' and housewives' ratings are equally numerous in the population and the researcher somehow achieved a probability sample, the mean ratings of Coke, Pepsi, and Shasta would be 5.0, 6.5, and 5.5, respectively. The mean results (and their ordering of the brands) would be shared by no individual. See the example in Table 2 of Lynch (1982:230) for elaboration. Thus, all of the action is in understanding the constructs (and indicators of them) that interact with the independent variables manipulated in our theory tests.

Consider the implications of these arguments for Moorman's (1996) award-winning quasi-experiment assessing the consumer and informational determinants of nutrition information-processing activities, as affected by the Nutrition Labeling and Education Act (NLEA). The NLEA mandated standard formats for provision of nutrition information. Moorman studied a sample of 1,000 consumers from balanced demographic, geographic, and site categories across 20 different product categories. These people were observed and then surveyed in a natural supermarket setting before and after NLEA.

I admire this article greatly, but not because I believe that the methods above confer high external validity on the findings, as so many other writers seem to believe. My

approach to external validity would point to the opportunity to learn by testing for interactions with product class, demographic, geographic, and site factors. An example of this is Moorman's (1996) assessment of differential effects of NLEA on nutrition information acquisition and comprehension for healthy and unhealthy products and her analyses splitting out effects on motivated versus unmotivated and skeptical versus unskeptical consumers. Without testing for these interactions, Moorman would risk significant aggregation fallacies and the some-none-all problem (Hutchinson et al. 1998), as in Lynch's (1982) cola example just cited. It was the tests for these interactions that make the research strong on the external validity dimension and not the representative sampling of real people, field methods, and so forth.

REALISM AND SOPHISTICATED CRITIQUES OF RESEARCH SETTINGS

Winer concurs with Wells's (1993) critique of the belief that "the laboratory represents the environment." That makes me nervous. Wells's point sounds a lot like the Ebbesen and Konecni (1980:38) view that I disputed in my 1982 article: "The really important truths are to be found in the real world rather than in laboratory simulations."

Such a view overlooks two important points. First, there is no reason to suspect that results from any single field setting are any more or less generalizable than those from any single laboratory setting. Dipboye and Flanagan (1979), writing about industrial and organizational psychology, note that the evidence is that findings from one field setting and from one lab setting are equally *unlikely* to generalize to a second field setting. Thus, the researcher who believes that her field study is necessarily high in external validity is naïve. Field researchers have just as much responsibility as lab researchers to engage in a thoughtful critique of the boundary conditions of their findings.

Second, if one's study is "unrealistic" on the level of some background factor that does not interact with the treatments, it has no effect on external validity. For this reason, Lynch (1982) argued that realism per se is irrelevant to external validity. If an experiment holds some background factor constant at an unrealistic level and if varying that background factor would have revealed a strong Treatment \times Background factor interaction, external validity is threatened. But the critic who seizes on some aspect of experimental procedure as unrealistic must explain his or her theory of why that variable should interact and show how the criticism explains the observed data (Lynch 1998). It is for this reason that I found Wells's (1993) arguments to be unpersuasive.

Particularly objectionable are calls for renouncing laboratory studies in favor of field studies that sacrifice internal validity in an attempt to maximize the ecological validity of the settings employed. (See Banaji and Crowder 1989 for an interesting critique of the "everyday memory" movement in cognitive psychology along these lines.) This is a case of chasing an illusory external validity benefit while ignoring the associated internal validity cost of moving to a field setting.² The researcher needs to think about realism at the construct level and not just at the level of surface features.

SUBJECTS: STUDENTS VERSUS REAL PEOPLE

Some years ago, the Marketing Science Institute (MSI) instituted a program in which researchers could very readily obtain funds to run their laboratory studies with "real" consumers instead of students. Although I believe MSI to have an overwhelmingly positive influence on the field of academic marketing, I thought that program was a clunker. It encouraged researchers (and MSI) to pay to recruit respondents with a high opportunity cost of time, with no clear analysis of what benefit might arise in terms of ability to understand the generality of findings.

I have the same reaction when Winer (1999) agrees with Wells (1993) that "real people" are inherently preferable to student subjects. Wells reiterated some fairly standard criticisms of using student subjects—implying that a homogeneous convenience sample of real people should be inherently superior. I disagree.

We need to raise the level of the debate from surface issues like whether the subjects are students or otherwise employed. Why would a single homogeneous group of church members or office clerks be any better? We should think at the level of constructs when we consider what biases might be engendered by relying on a sample of students. (See Sears 1986 for a thoughtful example.) The critic should specify (a) the construct on which the sample is atypical and (b) the case for why this construct might interact with the experimental treatment manipulations.

In some cases, a sample of students is likely to be relatively homogeneous on constructs that might matter—for example, involvement in a study of the processing of persuasive insurance ads. If the theoretical effects are ones that plausibly interact with involvement, that would threaten external validity. But if the phenomenon is one in which, theoretically, involvement should not matter (say, information processing phenomenon thought to operate automatically, below the level of conscious awareness), it would not be valid to criticize the use of (uninvolved) student subjects.

Note that the same restriction of range problem is equally possible when one selects a convenience sample of real people. In other cases, a critic's pet moderator variable might be just as varied in a sample of students as in a convenience or quota sample of real people. In this case, there is no advantage of real people over students—assuming that students' behavior falls within the domain of the theory. Better yet, though, would be to test two very different samples, treat this as a blocking factor in the analysis, and test for the interaction of sample source with the focal manipulations.

If a single convenience sample is to be used, there is one legitimate reason for preferring one type to another when both span only a narrow range on some likely moderator variable. Suppose that a researcher and her audience are likely to agree that a certain moderator variable is likely to interact with a treatment manipulation of focal interest in a project. Moreover, assume that the researcher believes that some levels of that background variable occur infrequently in the substantive systems of greatest interest to the researcher and her audience. In that case, one might prefer a homogeneous sample that has the "typical" levels of the background factor to one that has the atypical levels. For example, if one believes that prior knowledge interacts with some focal construct and one is interested in markets dominated by consumers with low levels of knowledge, one might prefer a study with a convenience sample of low-knowledge subjects to one with a convenience sample of high-knowledge subjects. This is a value judgment that priority should be given to more common levels if one can only illuminate one part of some complex response surface.

MAPPING FROM CONSTRUCTS TO OBSERVABLES

Winer (1999) suggests that the mission of a professor of marketing should be different from that of her counterpart working in a basic discipline like psychology or economics. Marketing professors serve constituencies of students and businesspeople who care deeply about the relevance of our research findings to particular substantive systems of interest to them. Winer argues that scholars who see themselves primarily as theory testers sell short our constituents "who are more interested in the real world than the laboratory world."

I agree. The development of useful theories is not only about getting it right in mapping the "structural model"—that is, the links among latent, unobservable constructs. It is also about mapping correctly from constructs to observables. Both sets of links are required if the audience is to be able to draw intelligent inferences about

the likely extrapolation of the findings reported to particular substantive systems that interest them.

Once one begins a thoughtful exercise of this nature, it becomes immediately apparent that real world versus lab is not a particularly useful basis for partitioning substantive systems just as student subjects versus real people is not a very insightful way to think about generalization across types of persons. One has to appeal to the construct level.

AN EXAMPLE OF A DISCUSSION SECTION FOLLOWING WINER (1999)

Winer (1999) uses his own past research to illustrate his ideal approach to external validity. Readers who know my characteristic modesty will no doubt be stunned that I find that I point to *my* past research to provide an appropriate model. The discussion section from Mitra and Lynch (1995) provides a good example of how an appropriate theoretical model should allow informed speculation about boundary conditions to the findings reported and about extrapolating to substantive domains in which different patterns might be expected on theoretical grounds. I discuss this example at some length to illustrate what ought to appear in Winer's proposed discussion section.

Mitra and Lynch (1995) were interested in the substantive problem of whether advertising increases or decreases consumer price elasticity. "Information" theories in economics held that advertising should increase price elasticity by making consumers more aware of the existence of substitutes. The "market power" school of thought held the opposite: advertising (artificially) differentiates brands that would otherwise be seen as close substitutes, conferring a degree of monopoly power on sellers.

Mitra and Lynch (1995) reconciled conflicting theories and data by positing that advertising affected price elasticity through its effects on two mediators, consideration set size and relative strength of preference. As shown in Figure 1, higher levels of advertising by all firms in a market increases consideration set size by a reminder function (Link 1) and increases relative strength of preference by providing differentiating information (Link 2). Stronger preferences for favorites causes consumers to drop less liked brands from their consideration sets (Link 3). Larger consideration sets are associated with greater (that is, more negative values of) price elasticity (Link 4). Greater relative strength of preference also directly reduces price elasticity (Link 5).

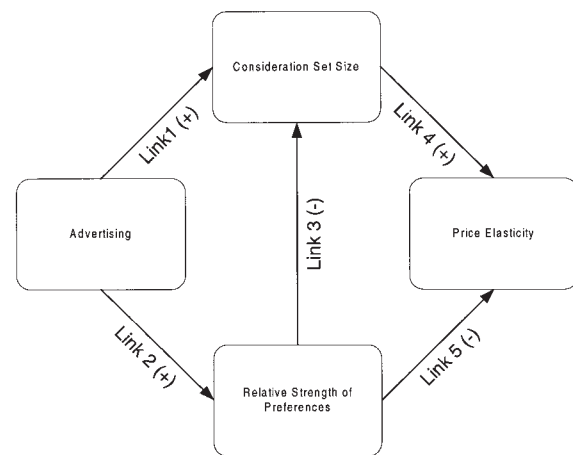
Mitra and Lynch (1995) conjectured that the net effect of advertising on price elasticity should be a function of the relative strength of these offsetting paths. We tested this in a laboratory experiment in which student subjects

spent real money to purchase Canadian candy bars. Our experiment varied independently two factors intended to alter the relative strength of the reminder path (Link 1) and the differentiating path (Link 2). We found that advertising had opposite effects depending on the strength of the reminder path. When the reminder function of advertising (Link 1) approached zero—because candy bar brands did not need to be remembered to be chosen—the net effect of advertising on price elasticity was negative. (The path through Link 2 \times Link 5 is negative, as is the path from Link 2 \times Link 3 \times Link 4.) When the reminder path was strong, advertising increased price sensitivity. The positive effect through Link 1 \times Link 4 outweighed the negative effects mediated through Link 2. We also showed that when the differentiating effect of advertising was weakened by providing less vivid information (making Path 2 closer to zero), the advertising was less prone to make consumers price insensitive.

Mitra and Lynch's (1995) discussion section, I think, follows the proposal by Winer (1999), modified by my friendly amendment. We mapped the strength of Link 1 into the degree to which the consumer relies on memory versus cues in the immediate purchase environment to form consideration sets. For example, Link 1 should be weaker in environments in which the alternative set is physically present and simple or when salespeople are pivotal to generating consideration sets. Link 1 should have a more important effect when the stimulus environment is highly complex, so one has to remember what one is looking for to consider brands—as in most grocery store environments and when time is short and motivation to search externally is low. This characterizes most mundane grocery shopping. The strength of Link 2 depends on the degree of prior familiarity with the brands advertised. The more consumers know, the less the power of advertising to differentiate further. For mature product categories, then, advertising by all firms in the market may still remind (Link 1) but fail to cause incremental differentiation (Link 2). Other arguments were made about real-world factors affecting the strength of Link 3. This kind of discussion permits a reader to conjecture in an informed way about likely effects of advertising on price elasticity in markets of particular interest to them.

While this article was going through the review process, we presented it at a conference, at which a discussant illustrated how *not* to think about external validity. He noted that it was unrealistic for subjects in our studies to have viewed ads for 12 candy bars twice in a given sitting under conditions of forced exposure and right before shopping. In his view, the findings were therefore ungeneralizable. Our approach was to respond at the level of constructs. Conceptually, the question is whether the

FIGURE 1
Effects of Advertising on Price Elasticity



SOURCE: Mitra and Lynch (1995:645).

findings are sensitive to the level of learning of the advertising message. Compared to, say, real markets for candy bars, our experiment included some facets that facilitated learning (short interval from learning to shopping) and some that inhibited it (interfering presentation for 12 competing brands, small number of total exposures). So it was not obvious from this critique that one should expect dissimilar findings in some real-world candy bar markets.

The same discussant questioned the realism of having U.S. subjects seeing ads for Canadian candy bars and making purchases of them. Should the results change if U.S. candy bars were used instead? Here, our answer was that we would *not* expect to replicate our findings. Our choice of unfamiliar Canadian candy bars was a deliberate attempt to maximize the strength of the differentiating effect of advertising via Link 2. With U.S. candy bars, we would expect our results to tip more toward findings consistent with the “information” view that advertising by all firms increases price elasticity.

It will consume precious journal pages to add Winer's (1999) suggested discussion analyzing external validity issues. What can we cut to make room? I propose deleting the typical limitations section, so often spent on boilerplate breast-beating about the limitations of using laboratory methods and nonrepresentative samples, especially students. If there is no speculation about specific constructs that might interact with the main findings, this is wasted space.

SCANNER DATA

I have spent much of this commentary discussing the first of Winer's (1999) proposals—to add to the discussion of every laboratory theory test a few paragraphs about matters of external validity. I will be much briefer in my comments about the second and third proposals, because my points follow from those made above.

Winer suggests that laboratory researchers seek external validity by testing for predicted effects using scanner panel data. A lot of interesting research has been generated from scanner panel data, but I do not see this as unique to the method. For the laboratory researcher working alone, there are large costs of learning how to handle the massive data sets generated—not unlike the learning curves for other specialized methodological expertise such as dealing with Internet log files, information integration methodology, ethnography, or the analysis of hemispheric brain activity. Moreover, for most behavioral theories, reliance on scanner panel data is likely to lead to omitted variable bias. The companies that operate household scanner panels prohibit surveying their panelists to ask attitudinal and cognitive questions that might be required to test most behavioral theories rigorously. Researchers who are accomplished in this method tend to seek out problems for which the method is appropriate and to ignore (equally interesting) problems for which the method would be seriously limited.

My interpretation is that Winer attributes to scanner panel analysis a special power to attain external validity. Dipboye and Flanagan (1979) argued that findings from single field settings have no special likelihood to be general; I would make the analogous claim that scanner analysis of a single packaged good like yogurt may well lack generality. If we start replicating the analysis across multiple categories and meta-analyzing the results, then we are really starting to add information to assess external validity (Lodish et al. 1995). But the same could be said about laboratory experiments with significant attention to stimulus replication (e.g., Veryzer and Hutchinson 1998). I can imagine reasons why I might advise the skilled laboratory researcher to drop her strength to spend a couple of years becoming versed in scanner data. Attaining higher external validity is not one of those reasons.

JOINT VENTURES

Winer (1999) makes a good point, though, in saying that laboratory leopards need not change their spots. They can partner in joint ventures with, say, a marketing scientist skilled in the analysis of panel data from real purchases. I endorse this proposal—cautiously. The idea that

multiple methods are desirable is mother's milk and apple pie. It is heresy to question the value added. But multiple methods help only if each method shores up some specific weakness the other has for studying the problem at hand.

Imagine that the Mitra and Lynch (1995) experiment was accompanied in the same article by one of two hypothetical follow-up studies. Study A is a second laboratory experiment that tests the conjecture that Mitra and Lynch's results would not replicate if subjects had already tasted the advertised products before they were exposed to advertising. Theoretically, more prior knowledge should weaken Link 2 (the differentiating effect of incremental advertising). This should cause reminder effects to play a bigger relative role in price elasticity. Subjects are randomly assigned to versions of Mitra and Lynch's experiment in which they do or do not taste the Canadian beers (rather than candy bars) prior to advertising exposure. Results replicate Mitra and Lynch when they have not tasted the beers beforehand but show a dominance of the reminder path over the differentiating path when the beers have been tasted.

Study B is a joint venture of Mitra and Lynch with an econometrician. It reports field data from vending machine sales consistent with the notion that higher industry spending on differentiating advertising is correlated with lower price elasticity. The study replicates the corresponding conditions from Mitra and Lynch's laboratory study, when the reminder effect of advertising should be unimportant. It makes no new theoretical points. Moreover, when taken by itself, it is equally consistent with a reverse causation account—firms spend more on advertising when their consumers are less price sensitive and, hence, more profitable to serve.

I would argue that follow-up Study A makes a bigger contribution than follow-up Study B (based on a joint venture). Moreover, a failure to replicate in Study B would not be particularly informative, as it could be due to a variety of theoretically relevant and irrelevant differences between that study and the Mitra and Lynch (1995) experiment. There are, however, cases in which joint ventures really contribute to the theorists' aims. I describe three of these below.

Intervention Falsification and Construct-to-Observable Mappings

When the goal is "intervention falsification" (Calder et al. 1981), the researcher may seek to demonstrate that it is possible to map from the constructs tested in the tight, laboratory study to a noisy real-world environment of particular interest. She is testing her assertions about the construct-to-observable mappings rather than adding any new insight on construct-to-construct relations. A

potentially confounded study that shows that the theoretically based intervention works, as implied by the theory, causes mild strengthening of belief in construct-to-construct parts of the theory (Brinberg, Lynch, and Sawyer 1992). It should have more dramatic effects on (presumably less certain) beliefs about construct-to-observable mappings. The Leclerc and Little (1997) article cited by Winer (1999) fits here.

Deliberate Sampling for Heterogeneity

Collaboration with a scholar versed in the analysis of observational data may allow a compelling instantiation of “deliberate sampling for heterogeneity” (Cook and Campbell 1979; Lynch 1982, 1983) when coupled with a laboratory study. In this mode, one is trying to demonstrate robustness rather than interactions with background factors. Two studies—or two blocks of the same study—differ widely in target classes of persons, settings, and times, although all are presumed to fall within the domain of the theory. The hope is that the findings will replicate. If they do not, there may be so many differences between lab and field that it is difficult to disentangle. Conceptually, deliberate sampling for heterogeneity might be followed in the lab without requiring the collaboration of a marketing scientist, but the demonstration of robustness may be especially strong from the kind of joint venture mode illustrated by Simonson and Winer (1992).

Combination of Constructs From Different Paradigms

Sometimes, collaboration with a scholar from another paradigm permits a fuller conceptualization of a substantive system of interest to researchers in both paradigms. The insight comes from viewing the same phenomenon at different levels of molarity. Steve Shugan affected my thinking greatly when he explained why he was a marketing scientist. He liked the idea that all the relevant players are modeled—for example, consumers, retailers, and manufacturers. All are adapting to the behavior of the others. I was struck by the difference from my world of experimental consumer research. There, seller behavior is exogenous, controlled by the researcher, as in the Mitra and Lynch (1995) example I have used to illustrate several of my points.

Recently, I have become involved in several research projects about the effects of electronic commerce on price sensitivity. Lynch and Ariely (1998) took on the argument that electronic commerce will lead to enhanced price sensitivity and ruinous price competition. We reported a laboratory experiment in which subjects spent their own money to buy wines from two competing online retailers

carrying some common and some unique wines. By changing the online interface, we varied orthogonally three different search costs in electronic shopping: the costs of price information, differentiating quality information, and comparing merchandise in two competing online retailers. We argued that the net effect of electronic commerce on price sensitivity is a matter of calibration of the relative strength of these three effects, so online commerce need not enhance price sensitivity and price competition. Without going into details, we found that the differentiating effects of lowering cost of search for quality information outweighed the effects of lowering the other two kinds of search costs.

A marketing scientist colleague, Preyas Desai, saw me present this work and commented that pricing was exogenous—the prices were set by the researchers. He asked the question of what would happen if seller pricing were endogenous. Subsequent conversations convinced us that the effect of lowered electronic search costs on prices is very subtle. We were able to derive conditions under which lowered search costs would have no effect, increase prices, or decrease prices, and we now have an embryonic experimental economics study designed to test our theorizing. I could never have done this alone.

I would make two points about this example. First, the value of the collaboration was not due primarily to our complementary methodological expertise; it was because Desai's paradigmatic knowledge allowed enrichment of the theory. Second, it should be admitted that some behavioral researchers are smart enough to be able to do excellent laboratory experiments and top-drawer modeling of the effects of economic institutions on both buyer and seller behavior—without joint ventures. Moorman's (1998) insightful analysis of the effects on sellers and buyers of the passage of the NLEA that mandated uniform provision of nutrition information provides an example. But most of us would do well to consider Winer's (1999) advice.

CONCLUSION

I strongly endorse Winer's call for sustained attention to matters of external validity. External validity and construct (nomological) validity are intimately linked. Like Winer, I am encouraged by recent trends toward multiple-experiment articles in which prior results are replicated with different subjects, operational definitions of variables, or procedures. Like Winer, I believe that the most informative “next” study is often the one that unconfounds “main” effects of the treatment manipulations in the past study from plausible interactions with background factors not part of the researchers' original theory. Where Winer

and I part company is that I believe that mundane realism has little scientific relevance and that the methods commonly thought to “enhance” external validity have no special power to do so. (See Lynch 1982, and especially 1983.)

External validity has less to do with what is in the method section of an article than with what is in the introduction, results, and discussion sections. It is the theory that motivates the study, the empirical assessment of interactions of the posited variables with background factors, interpretation of those findings in terms of more general constructs, and the mapping of the revised theory onto real-world observables that affect external validity. In all of this, we should be seeking to understand external validity at the level of constructs.

If theory-testing researchers choose important problems to study, their findings will naturally invite questions about their implications for real-world substantive systems. To extrapolate general knowledge about how constructs relate to some particular context (real world or otherwise) requires a step not taken in some theory-testing articles, as Winer (1999) has observed. One must conceptualize the mapping from the observables in the substantive system to a set of latent constructs in some nomological network. If Winer’s proposals are adopted with a focus of “external validity through theory” in mind, the science of marketing will prosper. We will be doing our jobs as scientists and business school professors.

ACKNOWLEDGMENTS

The author thanks Jonathan Levav, David Brinberg, and the editor, A. Parasuraman, for valuable comments on a prior draft.

NOTES

1. Cook and Campbell (1979:81-85) note that, even for applied research, internal validity is a precondition for external validity. Therefore, it is a mistake to follow procedures meant to enhance external validity that compromise internal validity.

2. My comments should not be interpreted as asserting a superiority of controlled laboratory experiments over other forms of research. All studies will prompt more belief shift, though, if they minimize the number of plausible alternative explanations for their results (Brinberg, Lynch, and Sawyer 1992).

REFERENCES

Banaji, Mahzarin R. and Robert G. Crowder. 1989. “The Bankruptcy of Everyday Memory.” *American Psychologist* 44 (September): 1185-1193.

Brinberg, David, John G. Lynch, Jr., and Alan G. Sawyer. 1992. “Hypothesized and Confounded Explanations in Theory Tests: A Bayes-

ian Analysis.” *Journal of Consumer Research* 19 (September): 139-154.

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.

Cook, Thomas and Donald T. Campbell. 1979. “Chapter 2: Validity.” In *Quasi-Experimentation: Design and Analysis Issues in a Field Setting*. Chicago: Rand McNally, 37-94.

Dipboye, Robert L. and Michael F. Flanagan. 1979. “Research Settings in Industrial and Organizational Psychology: Are Findings in the Field More Generalizable Than in the Laboratory?” *American Psychologist* 34:141-150.

Ebbesen, Ebbe B. and Vladimir J. Konecni. 1980. “On the External Validity of Decision Making Research: What Do We Know about Decisions in the Real World? In *Cognitive Processes in Choice and Decision Behavior*. Ed. Thomas Wallsetn. Hillsdale, NJ: Lawrence Erlbaum, 21-45.

Ferber, Robert. 1977. “Research by Convenience.” *Journal of Consumer Research* 4:57-58.

Hutchinson, J. Wesley, Wagner Kamakura, and John G. Lynch, Jr. 1998. “Unobserved Heterogeneity as an Alternative Explanation for ‘Reversal’ Effects in Behavioral Research.” Unpublished working paper, Wharton School, University of Pennsylvania.

Leclerc, France and John D. C. Little. 1997. “Can Advertising Copy Make FSI Coupons More Effective?” *Journal of Marketing Research* 34 (November): 473-484.

Lodish, Leonard, M. Abraham, S. Kalmensen, J. Livelsberger, B. Lubetkin, B. Richardson, and M. Stevens. 1995. “How TV Advertising Works: A Meta Analysis of 389 Real-World Split-Cable TV Advertising Experiments.” *Journal of Marketing Research* 32 (May): 125-139.

Lynch, John G., Jr. 1982. “On the External Validity of Experiments in Consumer Research.” *Journal of Consumer Research* 9 (December): 225-239. Erratum *Journal of Consumer Research* 9 (March): 455.

———. 1983. “The Role of External Validity in Theoretical Research.” *Journal of Consumer Research* 10 (June): 109-111.

———. 1998. “Presidential Address: Reviewing.” In *Advances in Consumer Research*. Eds. Joseph Alba and J. Wesley Hutchinson. Provo, UT: Association for Consumer Research, 1-6.

——— and Dan Ariely. 1998. “Electronic Shopping for Wine: Effects of Search Cost for Price and Quality Information on Consumer Price Sensitivity, Satisfaction With Merchandise, and Retention.” Unpublished working paper, Duke University.

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.

Mitra, Anusree and John G. Lynch, Jr. 1995. “Toward a Reconciliation of Market Power and Information Theories of Advertising Effects on Price Elasticity.” *Journal of Consumer Research* 21 (March): 644-659.

Moorman, Christine. 1996. “A Quasi Experiment to Assess the Consumer and Informational Determinants of Nutrition Information Processing Activities: The Case of the Nutrition Labeling and Education Act.” *Journal of Public Policy & Marketing* 15 (Spring): 28-44.

———. 1998. “Market-level Effects of Information: Competitive Responses and Consumer Dynamics.” *Journal of Marketing Research* 35 (February): 82-98.

Sears, David O. 1986. “College Sophomores in the Laboratory: Influences of a Narrow Base on Social Psychology’s View of Human Nature.” *Journal of Personality and Social Psychology* 51 (3): 515-530.

Simonson, Itamar 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.

Veryzer, Robert W. and J. Wesley Hutchinson. 1998. “The Influence of Unity and Prototypicality on Aesthetic Responses to New Product Designs.” *Journal of Consumer Research* 24 (March): 374-394.

Wells, William D. 1993. “Discovery-oriented Consumer Research.” *Journal of Consumer Research* 19 (March): 489-504.

Winer, Russell S. 1999. "Experimentation in the 21st Century: The Importance of External Validity." *Journal of the Academy of Marketing Science* 27 (3): 349-358.

ABOUT THE AUTHOR

John G. Lynch, Jr. is the Hanes Corporation Foundation Professor of Business Administration at Duke University. His research

and teaching interests are in consumer behavior, electronic commerce, and validity issues in research methodology. He is a past president of the Association for Consumer Research, past associate editor for the *Journal of Consumer Research*, and past associate editor and coeditor of the *Journal of Consumer Psychology*. He has been the recipient of the MSI/Paul Root Award at *Journal of Marketing*, the William O'Dell Award at *Journal of Marketing Research*, and has twice been the recipient of the *Journal of Consumer Research* best article award.