Integration of construct and external validity by means of proximal similarity: Implications for laboratory experiments in marketing

Edward F. McQuarrie*

Department of Marketing, Santa Clara University, Santa Clara, CA 95053, USA

Abstract

This paper applies the concept of proximal similarity to issues concerning the effective design of laboratory experiments in marketing. Proximal similarity focuses attention on the extent to which experimental procedures capture the distinctive characteristics of the marketing phenomenon under study. It requires a rethinking of the relationship between external validity and construct validity. The paper argues that proximal similarity is particularly crucial to an applied science, such as marketing, and concludes with suggestions on how the validity of laboratory experiments in marketing might be improved.

© 2003 Elsevier Science Inc. All rights reserved.

Keywords: Design of experiments; Construct validity; External validity; Mono-operation bias; Marketing science

1. Introduction

It is sometimes assumed that external validity has little to do with the design of effective laboratory experiments, whether in marketing or anywhere else. This paper offers a reconceptualization of the relationship between construct and external validity in terms of the idea of proximal similarity introduced by Campbell (1986). Experimental procedures commonly used in marketing are evaluated with respect to the achievement of proximal similarity. The paper then develops a set of recommendations as to how laboratory experiments in marketing must change if proximal similarity, and hence construct and external validity, is to be achieved.

The viewpoint offered in this paper opposes the position taken by Petty and Cacioppo (1996), who suggested that the "procedures of normal behavioral science" were perfectly adequate in marketing contexts. In turn, Petty and Cacioppo wrote in opposition to Wells (1993), who himself wrote in opposition to Calder et al. (1981), who in turn wished to oppose Ferber (1977) and others. The central issue in this long-running controversy concerns the role of external validity within a particular kind of research: laboratory experiments intended to advance theoretical knowledge. More specifically, debate has focused on the validity of using highly artificial stimuli and procedures when the goal is theoretical understanding of marketing phenomena. As discussed below, disagreements about the correct approach to the design of laboratory experiments rest ultimately on conflicting views about what a science of marketing entails, conflicts that go to the heart of the discipline's self-definition.

2. Conceptual interrelationship of construct and external validity

Current thinking about external validity in marketing science centers on the four-part validity scheme outlined in Cook and Campbell (1979). These authors expanded Campbell's (1957) original two-part validity scheme so as to differentiate external validity, construct validity, internal validity, and statistical conclusion validity. However, late in his life, Campbell (1986) wrote a paper in which he moved away from, and to some extent abandoned, the widely accepted Cook and Campbell (1979) synthesis. Essentially, Campbell argues that validity tradeoffs, as conventionally understood, may be neither wise nor necessary.

* Tel.: +1-408-554-6960; fax: +1-408-554-5056.
E-mail address: emcquarrie@scu.edu (E.F. McQuarrie).

0148-2963/ - see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S0148-2963(01)00298-3
2.1. Validity tradeoffs

The idea of a validity tradeoff rests on the assumption that specific research tools—i.e., the laboratory experiment—may be forced to minimize attention to one type of validity in order to excel at achieving some other type. The guiding analogy is to a zero-sum game, in which gains in one type of validity (e.g., internal validity) necessarily come at the expense of other types of validity (e.g., external validity). Most important, a belief in validity tradeoffs allows an investigator to affirm the importance of any given type of validity while continuing to ignore it in practice. Thus, no experimenter need argue that external validity is either unimportant or less important in order to justify a particular study design. Rather, external validity is someone else’s job, to be pursued at some other time, using some other research tool. Other priorities must govern when the goal is to design a laboratory experiment that will advance theoretical understanding of marketing phenomena.

For a validity tradeoff to be feasible, the two types of validity must be logically distinct or separable in principle. The value of the Campbell (1986) paper is twofold: on the one hand, it offers a radically different perspective on which pairs of validity types are logically distinct and which are not, while on the other, it provides a nuanced perspective on the kinds of separability that might exist. One surprising result concerns internal validity. Campbell argues that internal validity is logically distinct from construct validity (see Cronbach, 1982 for a similar view). Because laboratory experiments are typically justified as an attempt to advance theory, and because construct validity is an essential component of theoretical investigations, it follows that maximizing internal validity need not be an optimal strategy for enhancing the theoretical contribution of a laboratory experiment. In fact, it is possible that enhancements to internal validity may come at the cost of construct validity—a dubious tradeoff when theory is the goal.

Conversely, Campbell (1986) argues that construct validity and external validity are difficult to separate. The reintegration of construct and external validity highlights the extent to which Campbell’s own views evolved away from the Cook and Campbell (1979) synthesis (see Cook 1985, 1993 and Cook and Shadish, 1994 for an alternative evolution). If construct validity is not logically distinct from external validity but only factually distinct in certain particular circumstances (see below), it follows that it is perilous for an investigator to dismiss the relevance of external validity to the design of laboratory experiments when advancement of theoretical understanding is the goal.

If Campbell (1986) is correct, then the Cook and Campbell synthesis, at least as generally interpreted, cannot hold. What appeared to be acceptable tradeoffs, wherein attention to external validity is minimized, may actually be important threats to validity that undermine the scientific soundness of the work. We turn now to the exact nature of the relationship between construct and external validity, as it might be developed following Campbell (1986).

2.2. Notation

To begin, we need a precise way of describing how specific laboratory procedures implicate questions of internal, construct, and external validity (Fig. 1). Let $t$ stand for a treatment operation in some experiment, and let $o$ stand for an observed outcome. The treatment operation comprises all the concrete particulars of what was actually done—the manipulation as implemented at a given time and place. For example, suppose we encounter some research concerned with “source credibility and celebrity spokespersons.” At this level, we have as yet neither concrete particulars nor any treatment operation. Only when it is specified that the celebrity is Michael Jordan and the product is McDonald’s, that his endorsement will be encountered in print, that these printed words will appear on a slide projected in a dark room, and that this slide will be viewed by 20-year-old college students at a Big Ten University in a small Midwestern town in February do we begin to specify a treatment operation. The observed outcome is similarly concrete and particular: ratings on paired adjectives, arrayed across a white page in a specific pattern using a particular type face and point size, preceded by instructions of a certain sort, completed under a specific time limit, and so forth.

Given these definitions of treatment operation and observed outcome, we accept that a laboratory experiment has internal validity only to the extent that we are confident that this concrete particular treatment package ($t$) caused this omnibus outcome ($o$) in this time and place. For Campbell (1986), local, molar causal validity constitutes the entire extent of internal validity. By ‘local,’ Campbell means that no generalization can ever be said to possess internal validity; only a causal connection between a treatment and outcome located in a particular time and place can have or fail to have internal validity. By ‘molar,’ he means that it is the total treatment package, with all intended and unintended components, that is causally linked to the outcome exactly-as-measured. To apply a label to the treatment operation or outcome—to interpret them in terms of a theory of source credibility—is to raise questions of construct validity that are entirely distinct in logical terms.

In contrast to internal validity, which concerns the reliability, certainty, or trustworthiness of some specific finding of causal influence, both construct validity and external validity are held to involve some kind of generalization. Beyond this point, consensus breaks down (Cook, 1993). To clarify the issues, let $T$ and $O$ be theoretical constructs of interest (source credibility and brand attitude, say). A construct is a general idea that can never be fully captured by any single instance. Nevertheless, we accept that a laboratory experiment has construct validity only to
<table>
<thead>
<tr>
<th>Validity type</th>
<th>Focus of concern</th>
<th>Key question</th>
<th>Example threat</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal validity</td>
<td>Causality</td>
<td>Did this concrete, particular treatment ( t ) cause this observed outcome ( o ) in this time and place?</td>
<td>Both ( t_i ) and ( o_i ) were caused by a third variable</td>
<td>Involves no theoretical labeling and no generalization. Hence, renamed &quot;local, molar causal validity&quot; by Campbell (1986).</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Theoretical labeling</td>
<td>Is this ( t_i ) an instance of the theoretical construct of interest ( T )?</td>
<td>Confounding: ( t_i ) actually an instance of two theoretical constructs jointly (S and T)</td>
<td>To label ( t_i ) as an instance of ( T ) gives warrant to generalize results for ( t_i ) across the set ( { t_1, ..., t_k } ), consisting of all the treatment operations that index ( T ).</td>
</tr>
<tr>
<td>External validity (1)</td>
<td>Generalize across</td>
<td>Whether ( t_i ) will cause ( o_i ) for any ( t_i ) that is a member of the ( { t_1, ..., t_k } ) set</td>
<td>Only members of the ( { t_1, ..., t_k } ) subset—where theoretical construct ( S_a ) is present along with ( T )—cause ( o_i )</td>
<td>Systematic failure to generalize across other treatment operations indicates that ( t_i ) was mislabeled. This kind of external validity cannot be separated from construct validity.</td>
</tr>
<tr>
<td>External validity (2)</td>
<td>Generalize to</td>
<td>Whether a specific ( t_i ) drawn from ( { t_1, ..., t_k } ) will also cause ( o_i )</td>
<td>Failure to obtain the ( t_i ) ( \rightarrow ) ( o_i ) causal linkage</td>
<td>This kind of external validity can be separated from construct validity. Internal validity combined with construct validity provides only a warrant to generalize; any given attempt may fail for local, arbitrary reasons that do not call the construct validity of ( t_i ) into question.</td>
</tr>
</tbody>
</table>

Note: See text for further development of these points.

Explanation of notation:
- \( t_i \) = any concrete, particular treatment operation
- \( o_i \) = an outcome-as-measured
- \( T \) = a theoretical construct of interest
- \( S_a \) = a separate theoretical construct
- \( t_a \) = a treatment operation that reflects \( S_a \) as well as \( T \)

Fig. 1. Validity types after Campbell (1986).

The extent that we are confident that \( i \) and \( o \), the concrete and particular treatments and observed outcomes, do provide bona fide instances reflective of \( T \) and \( O \), the more general and broadly relevant theoretical constructs of interest.

Now, let \( \{ t_1, t_2, ..., t_k \} \) be a set of treatment operations, all of which index theoretical construct \( T \), varying only in terms of the theoretically irrelevant particulars that happen to be present in each case. To give an example of an irrelevant particular, there is no theory that suggests that reading print on a projected slide has any different effect than reading print on a magazine held in one's lap. Now, whenever we posit the existence of a theoretical construct \( T \), we may further assume that such a \( \{ t_1, ..., t_k \} \) set exists; the existence of such a set containing multiple operations follows directly from the fact that constructs cannot be fully captured by single concrete operations. The flip side of not being able to fully capture a general idea via a single operation is that for any general idea there must exist multiple operations capable of partially capturing it. This logic has important implications for the causal relations investigated in laboratory experiments. Whenever we demonstrate, with acceptable internal validity, that some particular construct valid treatment operation causes a certain effect, we have warrant to suppose that any other treatment operation drawn from the set \( \{ t_1, ..., t_k \} \) would also have a similar effect. Thus, if the only theoretical construct evident in the earlier example is Michael Jordan's celebrity status, we may assume that there is some set or population of celebrities to which he belongs, and we are warranted in supposing that the use of any other celebrity from that set would cause a similar pattern of observed outcomes.

The set \( \{ t_1, ..., t_k \} \) thus represents a population of treatment operations from which we sample whenever we conduct a particular experiment. This population of treatment operations will exhibit variance in how well any given operation instances the theoretical construct of interest. This means that sometimes when we sample a treatment operation, a large number of irrelevant particulars, or a particularly potent combination of particulars, will also be present and obscure the effect of the theoretical construct (Greenwood, 1989). In such cases, one happens to be dealing with a celebrity who, although a bona fide celebrity, is an atypical one whose atypicalities negate the theoretically expected advertising effect. In other cases, there will be few irrelevant particulars, or they will exert little influence, or their influence will cancel out, so that the expected celebrity effect is strong. The variance that characterizes the \( \{ t_1, ..., t_k \} \) set is why we speak of a warrant to generalize, meaning that we expect our generalization to hold when we sample anew from the \( \{ t_1, ..., t_k \} \) set that includes Michael Jordan, perhaps by a treatment operation involving Shaq O'Neal. However, this expectation is only that.
The assumption that underwrites any kind of generalization from concrete particular treatment operations is termed “proximal similarity” by Campbell (1986):

the presupposition that nature is ‘sticky,’ ‘viscous,’ proximally auto-correlated in space, time...with adjacent points more similar (as a rule) than non-adjacent ones...[hence] as scientists we generalize with most confidence to applications most similar to the setting of the original research (pp. 74–75).

The sticky, viscous, and proximally autocorrelated nature of the world justifies the expectation that if one member of a \( \{t_1...t_k\} \) set causes a certain outcome, a new sampling from that set will cause a similar outcome. Conversely, the principle of proximal similarity cautions us that the more distant our next treatment operation lies with respect to our initial investigation, the weaker our warrant for generalization.

With the aid of this notation, we can distinguish three cases with respect to the possible relationship between construct and external validity in a given experimental setting. Case 1 reflects the simplest possible case and is intended to show the inseparability of construct and external validity under certain conditions. Case 2 reflects the conventional wisdom as derived from the Cook and Campbell (1979) synthesis. Case 3 illuminates the difficulties with this synthesis, as developed by Campbell (1986).

2.3. Case 1: where construct and external validity cannot usefully be distinguished

When we expect the pattern of results obtained with Michael Jordan to generalize to Shaq O’Neal, are we asserting the construct validity or the external validity of the initial experiment? No distinction can be drawn between the two at this level (Campbell, 1986). If Michael Jordan is a member of a set (as he must be to index a theoretical construct), then results similar to those obtained for him can be expected to hold for other members of the set of operations reflecting that construct; we have no reason to suppose otherwise. It follows that any experiment that may be said to possess construct validity is guaranteed to have some degree of external validity—i.e., its findings can be expected to generalize to a range of other contexts (of greater or lesser extent). For if the pattern of observed outcomes associated with Michael Jordan could not be generalized, how could it be claimed that he indexed a general concept? Conversely, a zero level of external validity—a finding that could not be reproduced for any other treatment operation that we sample from the set—would automatically call into question the construct validity of the treatment operation examined. That is, our initial treatment operation was probably mislabeled—Michael Jordan was not really an instance of the set of celebrity treatments but of some other set.

2.4. Case 2: where it may be useful to distinguish construct and external validity

To see how construct and external validity can sometimes be separated, it is helpful to distinguish between two aspects of external validity. Once we have obtained an internally valid finding of causal influence using a specific treatment operation believed to have construct validity (e.g., Michael Jordan), we may wish to generalize this result in one of two ways, which were labeled by Cook and Campbell (1979, p. 71) as generalizing across vs. generalizing to. For example, our purpose in conducting an experiment using Michael Jordan might have been to determine whether an appeal from a celebrity—any celebrity—can induce brand switching for a frequently purchased convenience good. This would be an example of generalizing across the celebrity set in the context of a broad category of products. Alternatively, our purpose may have been to determine whether a black sports celebrity will be especially persuasive to white patrons of fast food restaurants in the Midwestern market during basketball season (an example of generalizing to a subset of cases). That is, we may wish either to predict events across substantially all treatment operations that implicate any instance of the celebrity set or we may intend the research to apply to some specific case or small subset of celebrity treatment operations.

This second sense corresponds to the narrow interpretation of external validity that has come to be associated with purely applied research (e.g., Petty and Cacioppo, 1996, p. 4). Thus, the goal of the typical industry copy test might be to measure the impact of a particular celebrity’s endorsement, presented as an animatic in a hotel room containing 100 adult homemakers in Columbus in February, with the intent of generalizing to results when the finished version of this commercial is placed on nationwide television in April—an even more specific example of generalizing to than the black sports celebrity and fast food example given just above. Calder et al. (1981) would label such copy tests “effects research.” Mook (1983) would term them “analogue research.”

Rarely or never would a study published in a scholarly journal be expected to possess a high degree of external validity in this restricted sense of generalizing to what will happen when a nationwide TV ad runs next April. Unfortunately, it seems to me that the (justified) dismissal by scholars of the need for external validity in the sense of generalizing to some specific real world instance has led, in the specific case of laboratory experiments, to a much more problematic neglect of the importance of external validity in the broader sense of the ability to generalize across.

To see why theorists must be concerned with generalizing across, let us assume for the sake of discussion that the goal of marketing science is to provide theoretical explanations of marketing phenomena. If so, then
whenever we do a laboratory experiment, the implicit claim is that the pattern of results obtained here in this artificial setting would also be obtained in many other settings found in the world, i.e., that we will be able to generalize across many instances of marketing phenomena. What we generalize are not the data values—not the one standard deviation increase in brand attitude toward McDonald’s that the Michael Jordan treatment creates in this particular experiment—but the theoretical interpretation of the pattern of results (Moock, 1983). If we fail again and again to get the same pattern of results in future studies when we use different treatment operations, then we have a failure of external validity, and this particular kind of failure of external validity will typically constitute a failure of construct validity as well. We probably did not correctly label our treatment operations in the initial laboratory experiment—either we measured something other than celebrity, we measured something more than celebrity, or we measured less than the full extent of the celebrity construct.

Now, consider a different kind of failure of external validity. Suppose our pattern of results fails to generalize when we devise a treatment operation using another celebrity—say, Madonna. Single failures of this kind may suggest limits to the external validity of our initial study but do not necessarily threaten its construct validity. We may have in Michael Jordan a construct valid treatment operation, and results with good external validity in the broad sense (i.e., applicable across a range of other celebrities), and nonetheless fail to correctly predict what will happen when we attempt to generalize to some very specific other celebrity (e.g., Madonna). A warrant to generalize is not a blanket guarantee. A theory may fail to hold up in specific instances for any number of local, arbitrary, and peculiar factors (e.g., Madonna’s pregnancy complicated her ‘celebrity’), without thereby undermining our confidence in the construct relations proposed by the theory. A good theory is generally (widely) applicable. It need not apply universally (completely, without exception) to continue to be regarded as a good theory.

To summarize, a failure to generalize to a specific concrete alternative treatment operation is a failure of external validity that is not per se a failure of construct validity. Here, the two types of validity are clearly distinguishable. Several failures to generalize to specific alternative treatments represent a kind of failure of external validity that also need not be a failure of construct validity (especially when there have also been successes). However, such multiple failures certainly invite further investigation to see whether they fit a pattern. Multiple failures to generalize to specific alternative treatments, where these do fit a theoretically interpretable pattern, are failures of external validity that are necessarily failures of construct validity (Lynch, 1983). In short, when external validity means generalizing to specific populations or settings, it can be distinguished from construct validity; when it means generalizing across theoretically interpretable patterns and settings, it cannot be distinguished from construct validity.

2.5. Case 3: where construct and external validity are distinct but mutually reinforcing

The preceding analysis has assumed a treatment operation that indexes a single theoretical construct (celebrity) in conjunction with some number of irrelevant parties (e.g., print on a slide). However, in the social sciences there are likely to be many cases where a treatment operation indexes more than one theoretical construct. Let \( \{t_1, t_2, \ldots, t_n, s_1, s_2, \ldots, s_m\} \) be a superset, containing all treatment operations that index other theoretical constructs (e.g., \( S_n, S_m, \ldots, S_k \)) in addition to the focal theoretical concept \( T \). For example, Michael Jordan is a black celebrity, and minority race (unlike the earlier print-slide vs. print-in-a-magazine contrast) is a variable located in various communication theories. Moreover, he is a sports celebrity, noted for his ability or expertise, and his reputation for being basketball crazy, and audience involvement with content of communications is also a theoretical variable. Although it remains the case that Michael Jordan is an instance of the theoretical construct ‘celebrity,’ the treatment operation that includes him may simultaneously be an instance of at least three other constructs: minority or expertise, and audience involvement.

The first consequence of this analysis is to place predictable bounds on the expected degree of external validity of our treatment operation. To recapitulate, demonstration of a causal effect for a treatment operation that indexes a construct warrants the expectation that a similar effect will hold for any other treatment operation sampled from the set that indexes that same construct. Hence, since the treatment operation under discussion appears simultaneously to index four constructs, our probability, in a Bayesian sense, that a similar outcome will be observed will be greatest for those new instances where all four constructs are again copresent—i.e., other cases where black sports celebrities communicate with diverse white audiences (cf. Brinberg et al., 1992). Our priors will be somewhat lower for instances where only two or three of these constructs are copresent and will be least for operations where only one construct is present. It must be emphasized that this is a question of probabilities only; the extent that any one of these constructs happens to be prepotent, empirical work may show that particular observed outcomes are generally present across a great number of instances of it, regardless of the copresence of the other theoretical constructs. It might even be possible to weigh these probabilities. Weights could be based on robustness—the frequency with which a construct, embodied in treatment, has had a measurable impact. Alternatively,
weights might be based on the magnitude of the impact the construct has had in previous implementations.

For example, ultimately, communicator celebrity may prove to be much more powerful than communicator race. Nonetheless, for any given laboratory experiment taken in isolation, the warrant for generalization is less the further the target context departs from the union of theoretical constructs indexed in the treatment operation actually implemented in the experiment. We cannot be as confident that our findings with Michael Jordan will generalize to Madonna (a non-basketball playing, white celebrity whose singing ability is sometimes doubted) as we could be with respect to, say, Shaq O'Neal.

It follows from this analysis that efforts to increase the construct validity of treatment operations will often act also to delineate the expected degree of external validity, i.e., will indicate exactly which superset of cases we have warrant to generalize across. Specifically, the techniques for increasing the construct validity of treatment operations turn out to be conceptually parallel to those applied in work on the construct validity of effects (e.g., the multitrait-multimethod approach to measurement). In both cases, multiple operations are key (Campbell and Fiske, 1959, O'Grady, 1982). If we can produce a similar pattern of outcomes with Larry Bird as well as Michael Jordan (neutralizing race), with the two of them presenting to an audience of sports-indifferent symphony and art students (neutralizing audience involvement), and with Roseanne and Madonna presenting to the original audience (neutralizing sports ability), then we will have a set of treatment operations whose only theoretical construct in common would appear to be celebrity. To the extent that celebrity is the focus of our research, then by means of these procedures we will have both considerably improved the construct validity of our treatment operations—our confidence that we have correctly labeled the Michael Jordan treatment operation—and we will have strengthened our warrant for generalizing across a wide variety of specific instances of celebrity—specifically, across the \( \{t_{1}, \ldots, t_{n}\} \) set (composed of celebrities of various races), the \( \{t_{11}, \ldots, t_{1k}\} \) set (composed of celebrities with varying degrees of ability/expertise), and the \( \{t_{y1, \ldots, k}\} \) set (composed of celebrities viewed by audiences varying in levels of involvement).

The preceding analysis indicates that the distinction between generalizing across and generalizing to, although helpful, is not absolute but a matter of degree. Our confidence that we have warrant to generalize across a wide variety of treatments increases with each successful generalization to a new and different treatment operation. Moreover, our initial confidence is higher if the first experiment includes multiple treatments (e.g., Madonna as well as Michael Jordan) and increases further when subsequent experiments employ yet other treatments whose distinctiveness is theoretically relevant. For example, we might investigate Magic Johnson after his AIDS diagnosis, or Dennis Rodham after an escape, in order to separate celebrity per se from positive celebrity (see Farley et al., 1998 for a related argument). If instead what Cook and Campbell (1979) call mono-operation bias is present—i.e., if we only use a single celebrity treatment operation—then we weaken our warrant to generalize across other kinds of operations. That is, introducing the possibility of a \( \{t_{1}, t_{2}, \ldots, t_{s}\} \) superset complicates matters, relative to the earlier formulation in terms of a \( \{t_{1}, \ldots, t_{k}\} \) set. On the assumption of a \( \{t_{1}, \ldots, k\} \) set, monooperation bias is not as worrisome. A strong case can be made that if internal validity has been secured, then reasonable warrant to generalize to other specific operations, and across the entire \( \{t_{1}, \ldots, k\} \) set, has been obtained. This warrant can always be strengthened by conducting a replication, but it does not require replication—we obtain the warrant by means of securing internal validity and assuming a "sticky, viscous, auto-correlated" world.

In circumstances where we cannot say with certainty whether we are dealing with a \( \{t_{1}, t_{2}, \ldots, t_{s}\} \) superset or a \( \{t_{1}, \ldots, t_{k}\} \) set, then monoo-operation bias becomes much more threatening. External validity is rendered more uncertain because we do not know whether the next treatment operation sampled will yield a similar effect. It will, if Michael Jordan and the new operation (Shaq O'Neal) belong to the same \( \{t_{1}, \ldots, k\} \) set; it may not, if the next operation (Roseanne) belongs instead to a \( \{t_{1}, \ldots, k\} \) set. Monoo-operation bias also undermines the construct validity of the experiment. If we do not know whether we have a \( \{t_{1}, \ldots, k\} \) set, then we also do not know whether Michael Jordan is best labeled as a celebrity, as a black celebrity, as an expert celebrity, or as a celebrity presenting to an involved audience.

If, as is common in laboratory experiments, we construct a treatment operation that resembles no real world instance whatsoever, this offers no inherent boost to external validity. A single utterly generic or artificial laboratory treatment, that resembles no specific real world case, does not gain in generalizability thereby; maximizing artificiality does not create a 'purified' index of the theoretical construct. A single artificial case gives us no warrant for generalizing across a wide range of instances beyond what we could get from using any single very concrete real world instance. All single treatment operations are equally fallible; only multiple operations (and multiple experiments) can directly enhance external validity in the sense of our confidence that we can generalize across cases, and our confidence that we know which kinds of cases we can generalize across, i.e., our confidence that have labeled the initial treatment operation correctly.

3. Proximal similarity and the design of experiments

3.1. Theoretical constructs and background variables

The distinction between \( \{t_{1}, \ldots, t_{s}\} \) sets and \( \{t_{1}, t_{2}, \ldots, t_{s}\} \) supersets provides a useful perspective on the debate

The set of background factors that could interact with treatments is infinite. Moreover, there is no a priori basis for even the most astute researcher to specify which of these factors will have an impact. Nor is there any logical way of even prioritizing these variables. . . in the absence of theory, there is, ipso facto, little reason to choose among possibilities (1982, pp. 241–242). Thus any attempt to include background factors . . . is necessarily ad hoc, leaving research open to all manner of extra-theoretical, intuitive biases that can only detract from the pursuit of theory (1983, p. 113).

In terms of the notation, Calder et al. are arguing that since we never have good theoretical grounds for exhaustively specifying the membership of the \( \{t_1, \ldots, t_n\} \) superset, we are better off assuming a \( \{t_1, \ldots, t_k\} \) set. Now, compare Campbell’s (1986) judgment on these matters: “Our intuitive expectations about what dimensions are relevant are theory-like, even if they are not formally theoretical” (p. 76). Campbell’s moderate view is important, because once the debate has been set up as theory or nothing, all background variables can be lumped together as of no moment.

A key assumption required to support the initial Calder et al. and later Sternthal et al. (1994) positions concerning background variables is that a strict separation in conceptual terms can be maintained between theoretical constructs and observed variables. This position can be traced back to a fundamental tenet of logical positivist and empiricist thought. However, beginning with Quine (1951), the notion that theory and observation terms could be sharply separated has been extensively debunked, and by the time of Suppe (1977), this idea had become untenable in mainstream philosophy of science (for more recent discussion of theory and observation terms, see Campbell, 1984, Hunt, 1994, and Shapere, 1984b). By assuming this separation, theory research is freed from having to worry about concrete particulars, and issues of proximal similarity become pertinent only within the subsequent stages of intervention and effects research, where generalizing to a narrow subset of cases is the focus (Calder et al., 1981). By contrast, for Campbell and his colleagues, major and very general theoretical constructs (e.g., source credibility) shade into minor and more localized ones (e.g., celebrity), which in turn shade into widely relevant nuisance variables (e.g., delayed impact), which shade into more narrowly pertinent background variables (e.g., print vs. video presentation modality). There is no separation in kind—central theoretical constructs and local nuisance variables are equally a matter of the interpretations placed on concrete, particular treatment operations and outcomes.

Moreover, Campbell suggests that in the case of complex human behaviors such as buying and selling, it may never be possible to construct an experiment that instantiates only a single theoretical construct. Rather, any concrete particular treatment operation can always be plausibly interpreted in terms of some number of background variables and nuisance theories in addition to the theoretical construct of interest. In terms of the notation introduced earlier, within the marketing domain, there probably is no such thing as a set of treatment operations \( \{t_1, t_2, \ldots, t_k\} \), which instance only one theoretical construct and which differ only in terms of irrelevant particulars. Rather, every realized treatment operation is actually a member of a \( \{t_1, t_2, \ldots, t_k\} \) superset indexing multiple constructs and variables. The important point is that if we never vary our treatment operations—if mono-operation bias is pervasive across an entire literature—then we cannot disentangle which effects are specific to the theoretical construct of interest, hence widely generalizable, and which only occur when certain background variables are conjoined.

To summarize, how we label our treatment operations—the stuff of construct validity—determines which subsets of other treatment operations we have warrant to generalize to and across—the stuff of external validity. Conversely, our successes and failures with respect to achieving external validity help us decide whether our initial labeling of the treatment operation was correct. Hence, under the assumption that treatment operations reflect multiple constructs, laboratory tests of theory must concern themselves with external validity and will naturally do so as part of vigorous efforts to secure construct validity. If with Campbell we also accept the utility of pretheoretical intuition, then marketing scientists who conduct laboratory experiments must be alert to background and nuisance variables in interpreting the results of these experiments, and this alertness can be expected to lead to multiple operationalizations of theoretical constructs tested over multiple experiments. In cases where it has not, the threat of mono-operation bias becomes very real, and a shadow is cast on both the construct and external validity of work in that area.

### 3.2. Treatments, constructs, and domains

The idea of mono-operation bias can be broadened by considering the notion of content domains. The concept of domain refers to any distinct area of investigation that requires a specialized set of concepts, tools, and procedures. A domain can be thought of as a cluster of proximally related phenomena more similar to each other than to more distal phenomena. For example, the domain of chemistry is distinguished from that of physics because understanding reactions at the molecular level requires a different approach than understanding interactions among subatomic particles (Hunt, 1991; Shapere, 1984a). The concept of a domain also applies within the human sciences, as in the distinction between psychology and sociology.

Let \( \tau \) be the set of domains in which some theoretical construct \( T \) is believed to have explanatory power. To return to our running example, suppose our goal was still to...
provide a theoretical explanation of advertising phenomena, but our initial treatment operation had instead comprised a videotaped Michael Milken, viewed by MBA students in a Finance class in New York City, challenging misconceptions about the perils of investing in collateralized mortgage obligations. By comparing this treatment to the earlier Michael Jordan example, we can integrate the three national elements. Thus, if celebrity were a universally applicable explanatory construct prepotent across all human domains, then both treatments can be considered members of a \{t_1, t_2\} set, and generalizing results from Michael Milken to Michael Jordan, or to any celebrity advertising appeal, would present no obstacle—the concept of domain does not apply. Similarly, if the union of the constructs celebrity + expertise + audience involvement was believed to be the crux, then both treatments would be \(t_n\) operations and generalization would again not be a problem. Whereas if more weight was placed on the differences between the two treatments with respect to communicator race, or media, or product type, then we would have a \(t_n\) vs. a \(t_N\) operation, and generalization would be constrained in the Bayesian fashion described earlier, but there would still be no difference in domain. It is when we argue that one treatment is an instance of education while the other is a matter of persuasion, and that the educational enterprise differs from the persuasion enterprise in a large number of respects that we do not know how to enumerate, that we would have a difference in domain. It is the large number of imperfectly understood differences, in conjunction with good reason to suppose that these differences matter (note the very different professional specializations and training regimens that have grown up around the two enterprises of education and persuasion), that suggests that the Michael Milken and Michael Jordan treatment operations reside in different domains. Even so, there should still be some warrant, however limited, for generalizing from the Michael Milken treatment to the Michael Jordan treatment, because both have construct validity as instances of a “celebrity spokesperson” or a credible source. However, because domains differ globally in countless ways, this warrant must be considered both weak and exceedingly difficult to estimate.

In sum, generalizing to one specific domain from research conducted in some other domain is fraught with all the difficulties that characterize the effort to generalize from a single treatment operation to some other treatment operation. Monodomain bias thus poses the same conceptual difficulties as monoooperation bias. If it is true that human behavior varies across domains, then it would be possible for, say, a highly stable body of results achieved in experimental social psychology with students making paper and pencil ratings of attitudes toward hypothetical dating partners to utterly fail to generalize to adult customers deciding how to allocate scarce dollars among brands to feed their families. When we want to explain phenomena in a specific domain with confidence (i.e., generalize to that domain), we must either conduct our research within that domain or have successfully generalized across enough heterogeneous domains in the past to have reasonable warrant for generalizing to the next new domain encountered.

No matter how many domains we have generalized across in previous work, we must still proceed with caution when we approach a new distal or dissimilar domain. At least, we should expect that the causal impact of our robust theoretical concepts might be conditioned on, or moderated by, background and nuisance variables distinctively characteristic of the new domain. In fact, until we have incorporated such variables into our methodologies, we cannot really claim to have empirically investigated the new domain.

Domains circumscribe the meaning of constructs. Had the concept of celebrity spokesperson been initially developed in the educational domain, then absent further investigation we could not be confident that it would apply in the persuasion domain. A change in domain changes the context of interpretation. Thus, whereas interpretation in terms of a theoretical construct acts to generalize a causal connection tested via a concrete treatment operation, acknowledgment that that causal connection has been tested in a particular domain acts to limit that generalization. The narrower the domain in which prior work has occurred, or the more distant the new domain from those that have hosted prior work, the greater the limitation on generalization. While in principle it is desirable to develop constructs that will be explanatory across many human domains, there is no guarantee that such generality can be achieved in the case of a specific construct or will hold when the focus shifts to a specific domain. We have the strongest warrant for generalization when we work within the same domain we are seeking to explain.

4. The domain specificity of marketing science

The principle of proximal similarity and the concept of domain would appear to be of particular relevance to someone who declares himself or herself to be a marketing scientist, as opposed to a “behavioral scientist.” That is, these ideas suggest the importance of insuring that even the most artificial laboratory experiment in marketing attempt to incorporate centrally located marketing phenomena in designing treatment operations. Thus, marketing scientists should prefer Michael Jordan endorsing McDonald's in an ad over Michael Milken endorsing collateralized mortgage obligations in an educational video, while behavioral scientists might remain indifferent. Unfortunately, it seems to me that a disinclination to accept the relevance of external validity in the design of experiments has instead led to an experimental tradition in marketing that by and large is not securely located within the marketing domain.

Note that a pure or basic behavioral scientist can stand unmoved by calls for proximal similarity. If the goal is a general theory of attitude change due to source credibility achieved via celebrity, then by definition such a scientist
will not be concerned with reproducing any specific domain of phenomena in laboratory experiments and will not care whether Michael Jordan or Michael Milken forms the basis of a treatment operation. As Petty and Cacioppo (1996) put it, such a scientist “is engaged in the examination of abstract and general hypotheses about human (consumer) behavior” (p. 4) and need only rely on “the procedures of normal behavioral science.” The concept of domain challenges the logic of such an approach because it raises doubt that a single set of concepts or procedures can be explanatory across all domains of human behavior.

Petty and Cacioppo (1996) go on to claim that the key distinction that governs the design of laboratory experiments, in marketing or elsewhere, is whether a person is engaged in... basic research on whether credibility enhances persuasion... or whether a person’s interests lie more in addressing specific and practical questions (e.g., applied research on whether Celebrity X is more effective than Celebrity Y in selling Product Z) (p. 4).

Both Preston (1985) and Wells (1993) would challenge Petty and Cacioppo’s characterization as a false polarity. Petty and Cacioppo’s distinction recalls Calder et al.’s discussion of background variables—either one studies abstract, general hypotheses, or descends to consulting; ignores background variables altogether, or becomes lost among an infinite number of alternatives. Such all-or-none contrasts should make us suspicious, suggesting that rhetorical discourse, and not scientific thinking, is on offer. This paper seeks a middle ground between the polarities posed by Petty and Cacioppo (1996). The test of Celebrity X vs. Celebrity Y for Product Z has no place in scholarship. Conversely, the logic of abstract, general hypotheses, followed to its conclusion, leaves no place for a marketing science.

The idea of a marketing science is more narrow and specific than the idea of a behavioral science; hence, if the best science seeks the most general hypotheses, marketing science would either be an inferior enterprise relative to behavioral science or simply redundant, unnecessary. Here, it might be objected that it is the investigation of marketing topics that constitutes marketing science as a distinct endeavor, i.e., that marketing science is best understood as behavioral science applied to marketing topics. However, this redefinition provides no relief. For if behavioral science has been successful in developing general hypotheses, should not these explain marketing topics and phenomena just as well as they explain the topics and phenomena on which they were initially derived? If the abstract, general hypotheses of Petty and Cacioppo really exist, then it would be redundant to investigate them again in marketing contexts—general hypotheses are supposed to allow us to “generalize across” topical areas.

The way out of this logic trap is to acknowledge as a discipline that we care deeply about marketing topics and are committed primarily to the scientific explanation of marketing phenomena and only secondarily to the practice of behavioral science. In other words, our goal is to “generalize to” marketing phenomena. It may turn out that marketing phenomena are behavioral phenomena plain and simple, so that success in the larger enterprise of explaining human behavior will automatically translate into success in our more local enterprise of explaining marketing phenomena. However, we cannot assume this outcome in advance. If we make a primary commitment to marketing topics, then we allow for either future outcome to occur—development over time of a distinct marketing science, optimized for explaining the specifics of the marketing domain, or a realization as research accumulates that marketing phenomena are simply one more domain where the basic laws of behavioral science provide the fundamental explanatory content. Conversely, a primary commitment to the procedures of normal behavioral science—which is to say, to the irrelevance of the distinction between the Michael Jordan and Michael Milken treatment operations—would make it impossible ever to discover whether marketing constitutes a distinct domain with its own scientific procedures and explanations. By analogy, biologists dare not be ignorant of chemistry and have much to learn from advances in chemistry, but experiments with nonliving chemical elements are not expected to provide the fundamental explanatory content for biology.

The concept of domain specificity helps to explain why so many laboratory experiments published in the marketing literature make use of brand stimuli. After all, is there a formal theory within behavioral science that specifies that the social object known as a brand—a name offering to be bought and sold—differs in any fundamental conceptual terms from the myriad other social objects previously studied by academic social psychologists? I know of none. Is there any theoretical reason why a core construct such as source credibility should operate differently in the case of brands as opposed to other social objects? No. Hence, if I read Petty and Cacioppo and Calder et al. correctly, when the goal of research is “the examination of abstract and general hypotheses,” a laboratory stimulus’ status as brand/not brand has to be considered a background variable of no special importance, making its inclusion or exclusion irrelevant to the contribution of the experiment to behavioral science.

Why then do experiments in marketing journals typically make use of brand stimuli? I believe it is precisely because most marketing scientists hold the pretheoretical intuition that brand behavior, embedded as it is in social relations constituted by buying and selling, may be a distinct domain of human behavior in which emergent phenomena not heretofore glimpsed by “normal behavioral science” may be found. That is why proximal similarity is more important to marketing science than to a pure or basic behavioral science—we are charged with generalizing our experimental
results to a specific domain rather than across all human domains. The meaning of a "general hypothesis" cannot be the same in marketing science—a domain-specific endeavor—as in general behavioral science.

The use of brand stimuli is a type case that demonstrates the importance and the necessity of employing pretheoretical intuitions based on considerations of proximal similarity in the design of marketing experiments. If our goal is to test "abstract, general hypotheses," it is not clear why we should care whether experiments contain brands or not. The fact that we do care suggests that the inclusion of brands actually functions as a domain marker, a reassurance that our research is in fact located within the marketing domain. This discussion is not meant to equate brand marketing with all of marketing nor to set up brand as a sine qua non for locating a study within the marketing domain. The concept of brand is simply a handy example of a domain marker. The thrust of this paper is that brandedness is but one example of the kind of domain marker that one might expect to find in scientific investigations dedicated to the explanation of marketing phenomena in particular. Section 4.1 offers an extended example of what domain specificity might entail in the case of one particular substantive topic area within marketing.

4.1. An example of domain specificity: advertising experiments

Advertising experiments offer a good opportunity for exploring the meaning of domain specificity, in part because there is a well-established literature in the behavioral sciences concerning communication and persuasion—the generic human behavioral phenomena of which advertising represents but one instance. Thus, the procedures for a "normal behavioral science" approach to laboratory studies of persuasion phenomena have been established at least since Hovland et al. (1953). Hence, it is straightforward to ask, What sort of experimental procedures might we expect to find in an experiment designed to be located squarely in the domain of advertising, but which, by contrast, would be either optional or unnecessary in a generic behavioral science investigation of persuasion? This amounts to a search for features that distinctively characterize advertising phenomena. Because these features are distinctive—not regularly found in other persuasion phenomena—normal behavioral science, with its drive toward abstract, general hypotheses, can freely ignore them in the design of persuasion experiments. Conversely, to say that these factors are characteristic means that a typical instance of the advertising phenomenon will evidence all or most of these factors. Hence, an experiment that reproduces these factors in its procedures can be expected to provide a stronger warrant for generalizing to the advertising domain.

With no attempt to be exhaustive, here are five examples of factors that would appear to be distinctively characteristic of advertising phenomena.

1. Advertisements are seldom the central focus of audience attention. Instead, it is the TV programming or print editorial in which the ad is embedded that primarily engages the audience (Allen and Madden, 1989; Kover, 1995).

2. The ultimate goal of advertising is to alter the choice behavior of the audience so as to enhance sales and profitability. By contrast, many other persuasion attempts are focused on shaping opinion, influencing a vote, affiliating with or joining a group, refraining from some unhealthy behavior, etc.

3. Ads must succeed in altering consumer brand choice in the face of competitive efforts running in the opposite direction. An experiment that presents a brand in isolation from any competitive brand sets up a choice between Brand A and not-Brand A (if choice is measured at all). By contrast, in advertising, the choice is mostly Brand A vs. Brand B (or C, or D...), with all brands having some opportunity to deliver messages to buyers.

4. Except for direct response or point-of-purchase advertising, ads must achieve their effect after a delay. In the case of most mass media advertising, the first opportunity to choose and buy a brand will occur at least a day after exposure.

5. Most advertisements will be exposed repeatedly. Although repetition is generally planned for in the case of advertising, it is not a necessary or typical feature of many other persuasion encounters.

The five factors just described are jointly characteristic of most mass media advertising efforts. Moreover, they would routinely be found in most field experiments involving advertising (e.g., Lodish et al., 1995). Conversely, it is easy to imagine constructing a bona fide persuasion experiment in the laboratory that would include none of these factors. Hence, including any one of these factors in an experiment represents a design choice with associated costs. A marketing scientist, under the argument of this paper, would be willing to accept these costs, in pursuit of proximal similarity and domain specificity, while a behavioral scientist need not.

What design choices have advertising scholars in fact elected over the years? A recent meta-analysis, from which the above list was drawn, examined hundreds of advertising experiments conducted over five decades (McQuarrie, 1998). It found that each of the five factors was ignored in 80–95% of the experiments. Moreover, later experiments (especially post-1980) were less likely to include these factors than experiments published earlier in the history of the marketing discipline. The meta-analysis suggests that in the specific case of advertising, marketing scholars have thus far not accepted the need to design experiments with domain specificity in mind.

Although I do not know of any similar analyses examining domain specificity in other areas of marketing, the logic of the advertising example can easily be extended to other marketing topics. Thus, experiments in retailing could be examined to see if they surround the respondent with possible choices while imposing time and budget constraints. A plethora of options would thus be consid-
ered to be a distinctive characteristic of retailing environments. Similarly, experiments concerned with fund raising by nonprofits might supply subjects with real money that they could either retain for their own benefit or donate in response to various experimental appeals (thus modeling donation as an action with real costs to oneself). In each case, we would be attempting to reproduce in the laboratory a distinctive characteristic of the marketing phenomenon in question, thereby achieving some degree of proximal similarity.

5. Toward a more domain-specific marketing science

The argument of this paper can be summarized as follows: (1) marketing phenomena—relations of buying and selling or exchange—constitute a distinct domain within human behavior; (2) as such, individual marketing phenomena exhibit distinctive characteristics or domain markers; (3) if a laboratory experiment is to provide a valid explanation of a marketing phenomenon, it must identify and strive to incorporate into its procedures as many of these domain markers as feasible. Proximal similarity is achieved by a disciplined search for domain markers and, once achieved, affords a combination of construct and external validity.

What would have to change if the argument of this paper were to become the accepted view? First, laboratory experiments designed to test theory would and should remain a staple of marketing science. Little would be gained by abandoning the laboratory in order to conduct field experiments exclusively. Although a single field experiment may have superior validity when we wish to generalize to some closely related instance, isolated field experiments provide no more warrant for generalizing across a wide number of cases than do isolated laboratory experiments (Locke, 1986). Moreover, laboratory experiments are markedly superior to field experiments from the standpoint of securing internal validity. Hence, the solution to the problems detailed in this paper must take the form of improvements in the design of laboratory experiments in marketing—abandonment is not the answer.

The first improvement would be to give domain specificity the same weight in experimental design now given to manipulation checks, to the construction of reliable measures, to controls against hypothesis guessing, and the like. For example, in an investigation of advertising, from the earliest stages, authors would assume the necessity of arranging for nonfocal attention to embedded stimuli. Only in special cases, carefully justified, would this element of the marketing domain be excluded. In turn, reviewers would begin to question the contribution of studies not clearly located in the marketing domain. Editors of marketing journals would begin to steer authors of such studies to other journals more suited to abstract behavioral science investigations.

A second consequence might be a shift in the choice of topics investigated. Marketing scientists might begin to seek out and focus on those marketing phenomena that have few parallels in other domains. For example, investigations of attitude theory might decline in frequency, while investigations of purchase behavior might increase. In addition, these purchase studies might focus especially on situations where multiple options are available, and the choice of one means giving up resources and foregoing the others. This shift in topics would be motivated by a desire to fulfill the promise of marketing as a science that can contribute insights into human behavior not otherwise obtainable.

6. Conclusion

Campbell's (1986) reconceptualization of the Cook and Campbell (1979) validity framework forces a reexamination of conventional wisdom in marketing concerning the design of laboratory experiments. The conventional (and comforting) assumption has been that validity tradeoffs are both necessary and unproblematic. Specifically, external validity has been regarded as something that can be deferred without threatening the theoretical contribution of the experiment. The reanalysis offered here of the interrelationship of construct and external validity, and the associated ideas of proximal similarity and domain specificity, reveals the pitfalls of this approach. Because external validity, in the sense of a capacity to generalize across, cannot be separated from the labeling of treatment operations, and because the theoretical interpretation of treatment operations hinges on how they are labeled, it follows that theory research cannot defer consideration of external validity. If we desire a theoretical understanding of marketing phenomena, then our experimental procedures, by virtue of proximal similarity and the incorporation of domain markers, must recognizably reproduce distinctive characteristics of marketing phenomena.

Two objections to this line of argument can be envisioned. These may be labeled “Have patience” and “It costs too much.” The first objection suggests that the deferral of external validity is a temporary phase in the development of the still-young discipline of marketing. That is, as time passes, we can reasonably expect the external validity of marketing experiments gradually to be filled in and extended. Unfortunately for this objection, both Preston (1985) and McQuarrie (1998) argue that the observable trend goes in the wrong direction. At least as far as the topic of advertising is concerned, experiments have grown less proximally similar over time.

The second objection is true up to a point. The specific revisions to experimental design suggested here will tend to increase the cost, in terms of both money and effort, of marketing experiments. We should recognize that resource constraints are very real in the case of theoretical marketing science. On the one hand, the kind of government funding evinced for medical science tends not to be available to


Quine WVO. Two dogmas of empiricism. Philos Rev 1951;60:20–43.


