

# Re-Inquiries

---

## The Perils of $N = 1$

WILLIAM D. WELLS\*

Research projects have three phases: (1) design, (2) execution, and (3) systematic exploration of range and limits. To achieve external credibility, positivistic and interpretive consumer researchers must pay close attention to phase 3.

---

Imagine the following: A Ph.D. candidate proposes an experiment on humor in advertising. In this experiment, a compliant sophomore is to look at a humorous advertisement for fictional brand A and a serious advertisement for fictional brand B. This respondent will then fill out a questionnaire that measures attitudes toward brands. The Ph.D. candidate asserts that if attitudes toward A are more favorable than attitudes toward B, the study will demonstrate that humorous advertising is more effective than serious advertising.

A typical committee would say something like this: "Although humor is a good topic, you must change your proposal. You must make sure that A was not better liked than B before exposure to the messages. You must justify your definition of effectiveness, and you must have a larger sample. You cannot generalize from an  $N$  of one." Of course the committee would be right. No serious investigator would generalize to an entire population from one convenient case.

So the candidate revises the proposal. Now a pretest will establish that A and B were equally liked beforehand, trusted literature will indicate that attitudes predict behavior, and  $N$  will equal 100 instead of one. The committee is much happier. The problem has been solved. Or has it?

Remember that the topic is humor in advertising. Humorous advertisements differ from each other in many ways. Some are attractive and some are unattractive. Some are

gentle. Some are clever. Some are funny and some are not so funny. Some poke fun at men, or women, or animals, and some do not. Some are for new brands and some are for established brands. Some are for funny products and some are for serious products. Some are for services. Some are in newspapers. Some are in magazines. Some are on the radio. Some are on television. In short, the category is heterogeneous.

Furthermore, homemade advertisements differ from real advertisements, and fictional brands differ from real brands. These differences affect relationships between attitudes and behavior. Still further, connections between attitudes and behavior are seldom very strong.

Finally, students differ from each other and from members of the wider population along dimensions that might affect results (Peterson 2000). Given this much chance for error, no responsible investigator would generalize to all humorous advertisements and all real consumers from the student's improved design. As the committee noted, one cannot make dependable generalizations to heterogeneous populations from one convenient case.

### NOT LIMITED TO ADVERTISING

The problem extends beyond advertising. Purchasing decisions assume many forms. They can be public or private, novel or habitual. They can be reasoned or impulsive, trivial or life altering. They can be selfish or generous, dictated or independent. They can demand elaborate tradeoffs among advantages and disadvantages or require little thought. Like humorous advertisements, purchasing decisions are heterogeneous.

All the differences within the category affect confidence in purchasing decisions and satisfaction with purchasing decisions. Given this heterogeneity, is it reasonable to draw general conclusions about confidence in purchasing deci-

---

\*William D. Wells is Mithun Land Grant Professor Emeritus in the School of Journalism and Mass Communication at the University of Minnesota (wells004@maroon.tc.umn.edu). Before joining the Minnesota faculty he served as executive vice president and director of marketing services at DDB Worldwide, an international marketing communications company. Before DDB, he served as professor of psychology and marketing in the Graduate School of Business, University of Chicago, and as professor of psychology at Rutgers University. He thanks the editor for this opportunity to present these comments.

sions from confidence in answers to general knowledge questions? Is it reasonable to draw general conclusions about satisfaction with purchasing decisions from accuracy of guesses in coin-toss games (Alba and Hutchinson 2000)?

Similarly, cultural differences have many consequences. They alter symbolic meanings and social relationships. They designate major differences in financial resources, linguistic frameworks, power structures, religious affiliations, retail outlets, communication media, and sources of prestige. Some cultural differences affect older consumers more (or less) than younger consumers, males more (or less) than females, well-educated consumers more (or less) than poorly educated consumers. In view of all this heterogeneity, is it reasonable to draw general conclusions about culture-motivated purchasing from interviews with undergraduates in California, Hong Kong, and Japan (Briley, Morris, and Simonson 2000)?

Examples could be multiplied indefinitely. Often overtly, always by implication, academic consumer researchers base general conclusions about heterogeneous categories on narrow convenience samples of stimuli, situations, products, or respondents (e.g., Ariely 2000; Biehal, Stephens, and Curlo 1992; Gardinal et al. 1993; Gorn et al. 1991; Gotlieb and Sarel 1991; Henthorne, La Tour, and Natarajan 1993; Lord, Burnkrant, and Unnava 2001; and Maheswaran, Mackie, and Chaiken 1992). These examples are not isolated aberrations. They illustrate accepted practice in the field.

Somehow, academic consumer researchers have convinced themselves that sampling does not matter. They generalize from one line drawing to all real advertisements in all real media, and from one fictional product to all real products and all real services. They generalize from one piece of music to all music, from one fear appeal to all fear appeals, from one depressing TV program to all depressing TV programs, from one deceptive manipulation to all communications, and from one laboratory setting to all decision sites. They generalize (at least by implication) from local samples of college sophomores and MBA candidates to all consumers in the culture or throughout the world.

If challenged, they claim their work is exploratory only, or they confess limitations, or they assert that they are doing basic science. None of these excuses is acceptable. Exploratory anticipates verification. Confessing limitations does not disable them. No science assumes zero variance. Like medical researchers, consumer researchers are not entitled to presume external validity. That honor must be earned.

## RANGE AND LIMITS

McGrath and Brinberg (1983) divided research into three phases: (1) design, (2) execution, and (3) systematic exploration of range and limits. Design and execution establish internal validity. Range and limits establish external validity.

Despite rationalizations to the contrary, no amount of internal validity can make up for lack of external validity. If internal validity could replace external validity, medical researchers would never move beyond white rats.

But academic consumer researchers have focused on in-

ternal validity and evaded external validity (McGrath and Brinberg 1983). As a direct consequence of this convenient custom, they do not attract serious attention from the outside world (Winer 1999).

Marketers, consumer advocates, and government officials need to know how real consumers make real decisions. They need good theory, that is, theory they can trust. Their careers depend on real returns from real investments. They treasure valid guidance and use all the information they can get.

But when they read the academic journals, they find study after study where  $N = 1$  convenient set of stimuli, situations, products, or respondents. For good reason, they do not trust that work.

To earn external credibility, academic consumer researchers must raise their standards. They must refuse to settle for convenience, and they must test the range and limits of the findings they produce.

This means systematic conceptual replications (Hunter 2001) that employ representative (not necessarily random) samples of stimuli, situations, products, and respondents. It also means re-inquiries (Mick 2001)—preferably but not exclusively by independent investigators—to test potential moderators. And it means critical analysis of treatments that did not seem to take.

## NOMOLOGICAL BRACKETING

In an early critique of academic overgeneralization, Brunswick (1947) advocated random sampling of stimuli to achieve representative design. While academic consumer researchers cannot draw probability samples of stimuli, situations, products, or respondents, they can and should draw purposive samples of prospective moderators.

For instance, if the reference finding is that a particular humorous advertisement is more effective than a matched serious advertisement, the investigator might confirm that finding with other kinds of humor, other products, or other media. If the reference finding came from student subjects, the follow-up might focus on more typical adults.

If the reference finding is that a particular one-sided comparative argument is more persuasive than its two-sided counterpart, the boundary tester might check other arguments, other topics, other presenters, other media, or other respondents.

In these follow-up investigations, the reference finding might hold, or it might not. If the reference finding holds, knowledge has been extended. If the reference finding does not hold (and the failure to replicate cannot be attributed to imperfect execution or unreliable measurement), an important boundary has been found.

## WHICH MODERATORS?

Of the many possible moderators, which should be considered? The answer to that question depends on the theoretical or practical implications of the moderators themselves. The decision rule would be focus on the moderator or moderators that seem most likely to interest theorists,

practitioners, or (preferably) both. This criterion provides low-cost relevance insurance. If the moderator is obviously important, the investigator has a ready-made answer to the perennial questions, *Who cares? So what?*

To repeat, this strategy is appealing because the researcher wins both ways. If the potential moderator makes no material difference (and the lack of difference cannot be charged to poor design or unreliable measurement), research scholars and those who need their findings have a more accurate picture of the stimuli, situations, objects, and respondents within which the reference outcome is likely to be correct. This contribution is an addition to knowledge in the most literal sense of that overused term.

If the moderator does make a reliable difference, research scholars and those who need their findings know something they did not know before. With that new knowledge, they are less liable to court disaster by generalizing to stimuli, situations, objects, or respondents for which the reference finding will be false.

### REVERSING THE FILE DRAWER EFFECT

When a study fails to find an expected relationship, present custom is to tuck it away in a file drawer and go on to something else. In the absence of deep interest in range and limits, this custom seems reasonable. If the project did not turn out as expected, why not try something that might work?

But hiding projects that fail to confirm predictions has two negative consequences. First, it weakens meta-analysis. If most of the studies that fail to reach statistical significance remain unpublished, too many of the studies that do reach statistical significance have gained acceptability by chance (Hunter 2001).

The second negative consequence of the file drawer effect is that forgotten files may hold useful information. A study that fails to find an expected relationship may have crossed a boundary that should be marked.

A relationship can fail to reach statistical significance because measurement was unreliable or invalid, because the sample was not large enough, or because the study has crossed a boundary beyond which the proposed relationship no longer holds. When measurement, design, and sample-size problems can be ruled out, it may be possible to discover the unforeseen alteration that caused the null result.

Perhaps the stimuli are different or are interpreted differently. Perhaps the context is substantially different from previous contexts. Perhaps the objects are different or have different meanings. Perhaps the respondents are different in unanticipated ways. Careful speculation will always produce potential explanations. Many of these explanations can be checked out.

Although this detective work will not always catch the culprit, it is more than worth the effort because when it is successful it marks stimuli, situations, products, or respondents that fall outside the nomological perimeter. With new boundary markers, we now have more dependable pictures of when, where, under what conditions, and among which

segments of the consumer population the reference finding can be expected to be correct.

Thus, close attention to range and limits encourages serious reanalysis of apparent failures. When this reanalysis is successful, it provides a double benefit. It makes the nomological network more dependable, and it rescues intellectual investments that otherwise would be lost.

### RECOMMENDATIONS

In academic settings, authors usually demote boundary questions to back pages of journal articles, where they evade the issue with pro forma calls for more research. But those calls almost never get answered. The premium that journal editors and other gatekeepers place on originality encourages the next investigator to start with unrelated research questions and end with equally equivocal results (Hunter 2001).

Thus, the discipline's current reward structure fosters risky incompleteness. If the discipline is to attain external credibility, that reward structure must be changed.

The first recommendation here is that journal editors and other gatekeepers give first priority to submissions that confront rather than evade external validity. The second recommendation is that journal editors and other gatekeepers solicit and publish studies that set boundary markers. In addition to encouraging efficient use of intellectual resources, that policy leads to outcomes that merit trust.

These modest recommendations are squarely in line with proposals advanced by Alba (2000), Brinberg and McGrath (1985), Ferber (1977), Jacoby (1976), Lehmann (1996), Lutz (1991), Lynch (1982), Sheth (1972), Shimp (1993), Simonson et al. (2001), Winer (1999), Zaltman (2000), and other scholars who have sought to upgrade the discipline. The common core of those proposals is that internal validity is necessary but not sufficient. Generalizability is never presumable. That claim always must be checked.

### APPLIES EQUALLY TO INTERPRETIVE INVESTIGATIONS

Although the present discussion springs from the positivist tradition, it applies equally to interpretive research. Whether the study is positivistic or interpretive, generalization must be earned.

Here is but one example. In a participant observation study of the meanings of Thanksgiving, Wallendorf and Arnould (1991) found that the prevailing mood in table conversation was cheerful and lighthearted. Even when talk turned to recollections of emergencies, the common theme was danger overcome by mutual support. The authors attributed this finding to the holiday's historic focus on the benefits of kinship and family ties.

An interpretive researcher bent on testing range and limits might ask whether this finding applies equally to other family gatherings such as Christmas and the Fourth of July. Or, because the original participant observers were mainly mainstream college students, an interpretive cross-checker might

ask whether the Thanksgiving mood is the same, for the same reasons, among less fortunate families, or among Hispanic-American families, or African-American families, or Native American families. Or a boundary-seeking scholar might ask whether the mood is the same, for the same reasons, during celebrations of the Chinese Moon Festival, the principal family-oriented harvest ritual in another part of the world. To the latter question, the answer is no (Chen and Wells 1999).

Recent re-inquiries by anthropologists (Wilk 2001) show that generalizations from participant observation and thick description are far from certain. In these re-inquiries, new studies of the same culture by different investigators produced dramatically different interpretations. Here, the limiting condition seems to have been the investigator. If that conclusion is correct, it demands new evidence from new investigators. Meanwhile, it warns against naive acceptance of either case.

### NEW AND HIGHER STANDARDS

Theorists and practitioners need to know how far conclusions can be stretched. They need to know when, where, under what conditions, and among what segments of the world's consumer population any particular finding can be expected to be valid. If academic consumer researchers want external credibility, they cannot leave corroboration up to others. They must not evade external validity because, in the appraisal systems of intelligent outsiders, external validity is all that really counts.

These new and higher standards would elevate the discipline. Instead of hiding behind assertions of theoretical legitimacy or interpretive immunity, academic consumer researchers would verify their work. Instead of filling the journals with "exploratory only," they would devote resources to responsible confirmation. Neighboring academic disciplines would respect this unique rigor. Government and industry would value findings they can trust.

We must not settle for pretend knowledge. We have chosen to investigate real problems. We must seek and find real answers. The keys to that achievement are in the hands of the discipline's gatekeepers. If academic consumer research is to become a trusted science, the incentives that now encourage risky incompleteness must be converted into incentives that encourage dependable delineation of the limits of the work.

[David Glen Mick served as editor for this article.]

### REFERENCES

- Alba, Joseph W. (2000), "Dimensions of Consumer Expertise . . . or Lack Thereof," in *Advances in Consumer Research*, Vol. 24, ed. Stephen J. Hoch and Robert J. Meyer, Provo, UT: Association for Consumer Research, 1-9.
- Alba, Joseph W. and J. Wesley Hutchinson (2000), "Knowledge Calibration: What Consumers Know and What They Think They Know," *Journal of Consumer Research*, 27 (September), 123-156.
- Ariely, Dan (2000), "Controlling the Information Flow: Effects on Consumers' Decision Making and Preferences," *Journal of Consumer Research*, 27 (September), 233-248.
- Biehal, Gabriel, Debra Stephens, and Eleonora Curlo (1992), "Attitude toward the Ad and Brand Choice," *Journal of Advertising*, 21 (September), 19-36.
- Briley, Donnel A., Michael W. Morris, and Itamar Simonson (2000), "Reasons as Carriers of Culture: Dynamic versus Dispositional Models of Cultural Influence on Decision Making," *Journal of Consumer Research*, 27, (September), 157-178.
- Brinberg, David and Joseph E. McGrath (1985), *Validity and the Research Process*, Beverley Hills, CA: Sage.
- Brunswick, Egon (1947), *Systematic and Representative Design of Psychological Experiments*, Berkeley: University of California Press.
- Chen, Qimei and William D. Wells (1999), "Melodies and Counterpoints: American Thanksgiving and the Chinese Moon Festival," in *Advances in Consumer Research*, Vol. 26, ed. Eric J. Arnould and Linda M. Scott, Provo, UT: Association for Consumer Research, 555-561.
- Ferber, Robert (1977), "Research by Convenience," *Journal of Consumer Research*, 4 (June) 57-58.
- Gardinal, Sarah Fisher, David W. Schumann, Ed Petkus, Jr., and Russell Smith (1993), "Processing and Retrieval of Inferences and Descriptive Advertising Information: The Effects of Message Elaboration," *Journal of Advertising*, 22 (March), 25-34.
- Gorn, Gerald J., Marvin E. Goldberg, Amitava Chattopadhyay, and David Litvack (1991), "Music and Information in Commercials: Their Effects with an Elderly Sample," *Journal of Advertising Research* 31 (October-November), 23-32.
- Gotlieb, Jerry B. and Dan Sarel (1991), "Comparative Advertising Effectiveness: The Role of Involvement and Source Credibility," *Journal of Advertising* 20 (1), 38-45.
- Henthorne, Tony L., Michael S. La Tour, and Rajan Natarajan (1993), "Fear Appeals in Print Advertising: An Analysis of Arousal and Ad Response," *Journal of Advertising*, 22 (June), 59-70.
- Hunter, John E. (2001), "The Desperate Need for Replications," *Journal of Consumer Research*, 28 (June), 149-158.
- Jacoby, Jacob (1976), "Consumer Research: Telling It Like It Is," in *Advances in Consumer Research*, Vol. 3, ed. Beverlee B. Anderson, Ann Arbor, MI: Association for Consumer Research, 1-11.
- Lehmann, Donald R. (1996), "Knowledge Generation and the Conventions of Consumer Research: A Study in Inconsistency," in *Advances in Consumer Research*, Vol. 23, ed. Kim P. Corfman and John G. Lynch, Provo, UT: Association for Consumer Research, 1-6.
- Lord, Kenneth R., Robert E. Burnkrant, and H. Rao Unnava (2001), "The Effects of Program-Induced Mood States on Memory for Commercial Information," *Journal of Current Issues and Research in Advertising*, 23 (Spring), 1-16.
- Lutz, Richard J. (1991), "Editorial," *Journal of Consumer Research*, 17 (March), vii-xiii.
- Lynch, John G. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (December), 225-239.
- Maheswaran, Durairaj, Diane M. Mackie, and Shelly Chaiken (1992), "Brand Name as a Heuristic Cue: The Effects of Task Importance and Expectancy Confirmation on Consumer Judgments," *Journal of Consumer Psychology*, 1 (4), 317-336.

- 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.
- Mick, David Glen (2001), "From the Editor," *Journal of Consumer Research*, 28 (June), iii-vii.
- Peterson, Robert A. (2000), "A Meta-Analysis of Variance Accounted for in Exploratory Factor Analysis," *Marketing Letters*, 11 (August), 261-275.
- Sheth, Jagdish N. (1972), "The Future of Buyer Behavior Theory," in *Proceedings of the Third Annual Conference of the Association for Consumer Research*, ed. M. Venkatesan, College Park, MD: Association for Consumer Research, 562-575.
- Shimp, Terence (1993), "Academic Appalachia and the Discipline of Consumer Research," in *Advances in Consumer Research*, Vol. 21, ed. Deborah Roedder John and Chris T. Allen, Provo, UT: Association for Consumer Research, 1-7.
- Simonson, Itamar, Ziv Carmon, Ravi Dhar, Aimee Drolet, and Stephen M. Nowlis (2001), "Consumer Research: In Search of Identity," in *Annual Review of Psychology*, Vol. 52, ed. Susan T. Fiske, Daniel L. Schacter, and Carolyn Zahn-Waxler, Palo Alto, CA: Annual Reviews, 249-275.
- Wallendorf, Melanie and Eric J. Arnould (1991), "'We Gather Together': Consumption Rituals of Thanksgiving Day," *Journal of Consumer Research*, 18 (June), 13-31.
- Wilk, Richard R. (2001), "The Impossibility and Necessity of Re-Inquiry: Finding Middle Ground in Social Science," *Journal of Consumer Research*, 28 (September), 308-312.
- Winer, Russell S. (1999), "Experimentation in the Twenty-First Century: The Importance of External Validity," *Journal of the Academy of Marketing Science*, 27 (3), 349-358.
- Zaltman, Gerald (2000), "Consumer Researchers: Take a Hike," *Journal of Consumer Research*, 26 (March), 423-428.