In Defense of External Invalidity

Douglas G. Mook *University of Virginia*

ABSTRACT: Many psychological investigations are accused of "failure to generalize to the real world" because of sample bias or artificiality of setting. It is argued in this article that such "generalizations" often are not intended. Rather than making predictions about the real world from the laboratory, we may test predictions that specify what ought to happen in the lab. We may regard even "artificial" findings as interesting because they show what can occur, even if it rarely does. Or, where we do make generalizations, they may have added force because of artificiality of sample or setting. A misplaced preoccupation with external validity can lead us to dismiss good research for which generalization to real life is not intended or meaningful.

> The greatest weakness of laboratory experiments lies in their artificiality. Social processes observed to occur within a laboratory setting might not necessarily occur within more natural social settings.
>
> —Babbie, 1975, p. 254

> In order to behave like scientists we must construct situations in which our subjects . . . can behave as little like human beings as possible and we do this in order to allow ourselves to make statements about the nature of their humanity.
>
> —Bannister, 1966, p. 24

Experimental psychologists frequently have to listen to remarks like these. And one who has taught courses in research methods and experimental psychology, as I have for the past several years, has probably had no problem in alerting students to the "artificiality" of research settings. Students, like laypersons (and not a few social scientists for that matter), come to us quite prepared to point out the remoteness of our experimental chambers, our preoccupation with rats and college sophomores, and the comic-opera "reactivity" of our shock generators, electrode paste, and judgments of lengths of line segments on white paper.

They see all this. My problem has been not to alert them to these considerations, but to break their habit of dismissing well-done, meaningful, informative research on grounds of "artificiality."

The task has become a bit harder over the last few years because a full-fledged "purr" word has gained currency: *external validity.* Articles and monographs have been written about its proper nurture, and checklists of specific threats to its well-being are now appearing in textbooks. Studies unescorted by it are afflicted by—what else?—*external invalidity.* That phrase has a lovely mouth-filling resonance to it, and there is, to be sure, a certain poetic justice in our being attacked with our own jargon.

Warm Fuzzies and Cold Creepies

The trouble is that, like most "purr" and "snarl" words, the phrases *external validity* and *external invalidity* can serve as serious barriers to thought. Obviously, any kind of validity is a warm, fuzzy Good Thing; and just as obviously, any kind of invalidity must be a cold, creepy Bad Thing. Who could doubt it?

It seems to me that these phrases trapped even their originators, in just that way. Campbell and Stanley (1967) introduce the concept thus: "*External validity* asks the question of *generalizability:* To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (p. 5). Fair enough. External validity is not an automatic desideratum; it *asks a question.* It invites us to think about the prior questions: To what populations, settings, and so on, do we *want* the effect to be generalized? Do we want to generalize it at all?

But their next sentence is: "Both types of criteria are obviously important. . ." And ". . . the selection of designs strong in both types of validity is obviously our ideal" (Campbell & Stanley, 1967, p. 5).

I intend to argue that this is simply wrong. If it sounds plausible, it is because the word *validity* has given it a warm coat of downy fuzz. Who wants to be invalid—internally, externally, or in any other way? One might as well ask for acne. In a way, I wish the authors had stayed with the term *generalizability,* precisely because it does not sound nearly so good. It would then be easier to remember that we are not dealing with a criterion, like clear skin, but with a question, like "How can we get this sofa down the stairs?" One asks that question if, and only if, moving the sofa is what one wants to do.

But *generalizability* is not quite right either. The question of external validity is not the same as the question of generalizability. Even an experiment

that is clearly "applicable to the real world," perhaps because it was conducted there (e.g., Bickman's, 1974, studies of obedience on the street corner), will have *some* limits to its generalizability. Cultural, historical, and age-group limits will surely be present; but these are unknown and no single study can discover them all. Their determination is empirical.

The external-validity question is a special case. It comes to this: Are the sample, the setting, and the manipulation so artificial that the class of "target" real-life situations to which the results can be generalized is likely to be trivially small? If so, the experiment lacks external validity. But that argument still begs the question I wish to raise here: Is such generalization our intent? Is it what we want to do? Not always.

The Agricultural Model

These baleful remarks about external validity (EV) are not quite fair to its originators. In defining the concept, they had a particular kind of research in mind, and it was the kind in which the problem of EV is meaningful and important.

These are the applied experiments. Campbell and Stanley (1967) had in mind the kind of investigation that is designed to evaluate a new teaching procedure or the effects of an "enrichment" program on the culturally deprived. For that matter, the research context in which sampling theory was developed in its modern form—agricultural research—has a similar purpose. The experimental setting resembles, or is a special case of, a real-life setting in which one wants to know what to do. Does this fertilizer (or this pedagogical device) promote growth in this kind of crop (or this kind of child)? If one finds a significant improvement in the experimental subjects as compared with the controls, one predicts that implementation of a similar manipulation, in a similar setting with similar subjects, will be of benefit on a larger scale.

That kind of argument does assume that one's experimental manipulation represents the broader-scale implementation and that one's subjects and settings represent their target populations. Indeed, part of the thrust of the EV concept is that we have been concerned only with subject representativeness and not enough with representativeness of the settings and manipulations we have sampled in doing experiments.

Deese (1972), for example, has taken us to task for this neglect:

> Some particular set of conditions in an experiment is generally taken to be representative of all possible conditions of a similar type. . . . In the investigation of altruism, situations are devised to permit people to make altruistic choices. Usually a single situation provides the setting for the experimental testing. . . . [the experimenter] will allow that one particular situation to stand for the unspecified circumstances in which an individual could be altruistic. . . . the social psychologist as experimenter is content to let a particular situation stand for an indefinite range of possible testing situations in a vague and unspecified way. (pp. 59–60)

It comes down to this: The experimenter is generalizing on the basis of a small and biased sample, not of subjects (though probably those too), but of settings and manipulations.[1]

The entire argument rests, however, on an applied, or what I call an "agricultural," conception of the aims of research. The assumption is that the experiment is *intended* to be generalized to similar subjects, manipulations, and settings. If this is so, then the broader the generalizations one can make, the more real-world occurrences one can predict from one's findings and the more one has learned about the real world from them. However, it may not be so. There are experiments—very many of them—that do not have such generalization as their aim.

This is not to deny that we have talked nonsense on occasion. We have. Sweeping generalizations about "altruism," or "anxiety," or "honesty" have been made on evidence that does not begin to support them, and for the reasons Deese gives. But let it also be said that in many such cases, we have seemed to talk nonsense only because our critics, or we ourselves, have assumed that the "agricultural" goal of generalization is part of our intent.

But in many (perhaps most) of the experiments Deese has in mind, the logic goes in a different direction. We are not *making* generalizations, but *testing* them. To show what a difference this makes, let me turn to an example.

A Case Study of a Flat Flunk

Surely one of the experiments that has had permanent impact on our thinking is the study of "mother love" in rhesus monkeys, elegantly conducted by Harlow. His wire mothers and terry-cloth mothers are permanent additions to our vocabulary of classic manipulations. And his finding that con-

I thank James E. Deese and Wayne Shebilske for their comments on an earlier version of this article.

Requests for reprints should be sent to Douglas G. Mook, Department of Psychology, University of Virginia, Charlottesville, Virginia 22901.

[1] In fairness, Deese goes on to make a distinction much like the one I intend here. "If the theory and observations are explicitly related to one another through some rigorous logical process, then the sampling of conditions may become completely unnecessary" (p. 60). I agree. "But a theory having such power is almost never found in psychology" (p. 61). I disagree, not because I think our theories are all that powerful, but because I do not think all that much power is required for what we are usually trying to do.

tact comfort was a powerful determinant of "attachment," whereas nutrition was small potatoes, was a massive spike in the coffin of the moribund, but still wriggling, drive-reduction theories of the 1950s.

As a case study, let us see how the Harlow wire- and cloth-mother experiment stands up to the criteria of EV.

The original discussion of EV by Campbell and Stanley (1967) reveals that the experimental investigation they had in mind was a rather complex mixed design with pretests, a treatment imposed or withheld (the independent variable), and a posttest. Since Harlow's experiment does not fit this mold, the first two of their "threats to external validity" do not arise at all: pretest effects on responsiveness and multiple-treatment interference.

The other two threats on their list do arise in Harlow's case. First, "there remains the possibility that the effects . . . hold only for that unique population from which the . . . [subjects were] selected" (Campbell & Stanley, 1967, p. 19). More generally, this is the problem of sampling bias, and it raises the spectre of an unrepresentative sample. Of course, as every student knows, the way to combat the problem (and never mind that nobody does it) is to select a random sample from the population of interest.

Were Harlow's baby monkeys representative of the population of monkeys in general? Obviously not; they were born in captivity and then orphaned besides. Well, were they a representative sample of the population of lab-born, orphaned monkeys? There was no attempt at all to make them so. It must be concluded that Harlow's sampling procedures fell far short of the ideal.

Second, we have the undeniable fact of the "patent artificiality of the experimental setting" (Campbell & Stanley, 1967, p. 20). Campbell and Stanley go on to discuss the problems posed by the subjects' knowledge that they are in an experiment and by what we now call "demand characteristics." But the problem can be generalized again: How do we know that what the subjects do in this artificial setting is what they would do in a more natural one? Solutions have involved hiding from the subjects the fact that they are subjects; moving from a laboratory to a field setting; and, going further, trying for a "representative sample" of the field settings themselves (e.g., Brunswik, 1955).

What then of Harlow's work? One does not know whether his subjects knew they were in an experiment; certainly there is every chance that they experienced "expectations of the unusual, with wonder and active puzzling" (Campbell & Stanley, 1967, p. 21). In short, they must have been cautious, bewildered, reactive baby monkeys indeed. And what of the representativeness of the setting? Real monkeys do not live within walls. They do not encounter mother figures made of wire mesh, with rubber nipples; nor is the advent of a terry-cloth cylinder, warmed by a light bulb, a part of their natural lifestyle. What can this contrived situation possibly tell us about how monkeys with natural upbringing would behave in a natural setting?

On the face of it, the verdict must be a flat flunk. On every criterion of EV that applies at all, we find Harlow's experiment either manifestly deficient or simply unevaluable. And yet our tendency is to respond to this critique with a resounding "So what?" And I think we are quite right to so respond.

Why? Because using the lab results to make generalizations about real-world behavior was no part of Harlow's intention. It was not what he was trying to do. That being the case, the concept of EV simply does not arise—except in an indirect and remote sense to be clarified shortly.

Harlow did not conclude, "Wild monkeys in the jungle probably would choose terry-cloth over wire mothers, too, if offered the choice." First, it would be a moot conclusion, since that simply is not going to happen. Second, who cares whether they would or not? The generalization would be trivial even if true. What Harlow did conclude was that the hunger-reduction interpretation of mother love would not work. If anything about his experiment has external validity, it is this theoretical point, not the findings themselves. And to see whether the theoretical conclusion is valid, we extend the experiments or test predictions based on theory.[2] We do not dismiss the findings and go back to do the experiment "properly," in the jungle with a random sample of baby monkeys.

The distinction between generality of findings and generality of theoretical conclusions underscores what seems to me the most important source of confusion in all this, which is the assumption that the purpose of collecting data in the laboratory is to *predict real-life behavior in the real world.* Of course, there are times when that is what we are trying to do, and there are times when it is not. When it is, then the problem of EV confronts us, full force. When it is not, then the problem of EV is either meaningless or trivial, and a misplaced preoccupation with it can seriously distort our evaluation of the research.

But if we are not using our experiments to predict real-life behavior, what are we using them for? Why else do an experiment?

[2] The term *theory* is used loosely to mean, not a strict deductive system, but a conclusion on which different findings converge. Harlow's demonstration draws much of its force from the context of other findings (by Ainsworth, Bowlby, Spitz, and others) with which it articulates.

There are a number of other things we may be doing. First, we may be asking whether something *can* happen, rather than whether it typically *does* happen. Second, our prediction may be in the other direction; it may specify something that ought to happen *in the lab,* and so we go to the lab to see whether it does. Third, we may demonstrate the power of a phenomenon by showing that it happens even under unnatural conditions that ought to preclude it. Finally, we may use the lab to produce conditions that have no counterpart in real life at all, so that the concept of "generalizing to the real world" has no meaning. But even where findings cannot possibly generalize and are not supposed to, they can contribute to an understanding of the processes going on. Once again, it is that understanding which has external validity (if it does)—not the findings themselves, much less the setting and the sample. And this implies in turn that we cannot assess that kind of validity by examining the experiment itself.

Alternatives to Generalization

"What Can" Versus "What Does"

"Person perception studies using photographs or brief exposure of the stimulus person have commonly found that spectacles, lipstick and untidy hair have a great effect on judgments of intelligence and other traits. It is suggested . . . that these results are probably exaggerations of any effect that might occur when more information about a person is available" (Argyle, 1969, p. 19). Later in the same text, Argyle gives a specific example: "Argyle and McHenry found that targeted persons were judged as 13 points of IQ more intelligent when wearing spectacles and when seen for 15 seconds; however, if they were seen during 5 minutes of conversation spectacles made no difference" (p. 135).

Argyle (1969) offers these data as an example of how "the results [of an independent variable studied in isolation] may be exaggerated" (p. 19). Exaggerated with respect to what? With respect to what "really" goes on in the world of affairs. It is clear that on these grounds, Argyle takes the 5-minute study, in which glasses made no difference, more seriously than the 15-second study, in which they did.

Now from an "applied" perspective, there is no question that Argyle is right. Suppose that only the 15-second results were known; and suppose that on the basis of them, employment counselors began advising their students to wear glasses or sales executives began requiring their salespeople to do so. The result would be a great deal of wasted time, and all because of an "exaggerated effect," or what I have called an "inflated variable" (Mook, 1982). Powerful in the laboratory (13 IQ points is a lot!), eyeglasses are a trivial guide to a person's intelligence and are treated as such when more information is available.

On the other hand, is it not worth knowing that such a bias *can* occur, even under restricted conditions? Does it imply an implicit "theory" or set of "heuristics" that we carry about with us? If so, where do they come from?

There are some intriguing issues here. Why should the person's wearing eyeglasses affect our judgments of his or her intelligence under any conditions whatever? As a pure guess, I would hazard the following: Maybe we believe that (a) intelligent people read more than less intelligent ones, and (b) that reading leads to visual problems, wherefore (c) the more intelligent are more likely to need glasses. If that is how the argument runs, then it is an instance of how our person perceptions are influenced by causal "schemata" (Nisbett & Ross, 1980)—even where at least one step in the theoretical sequence ([b] above) is, as far as we know, simply false.

Looked at in that way, the difference between the 15-second and the 5-minute condition is itself worth investigating further (as it would not be if the latter simply "invalidated" the former). If we are so ready to abandon a rather silly causal theory in the light of more data, why are some other causal theories, many of them even sillier, so fiercely resistant to change?

The point is that in thinking about the matter this way, we are taking the results strictly as we find them. The fact that eyeglasses *can* influence our judgments of intelligence, though it may be quite devoid of real-world application, surely says something about us as judges. If we look just at that, then the issue of external validity does not arise. We are no longer concerned with generalizing from the lab to the real world. The lab (qua lab) has led us to ask questions that might not otherwise occur to us. Surely that alone makes the research more than a sterile intellectual exercise.

Predicting From and Predicting To

The next case study has a special place in my heart. It is one of the things that led directly to this article, which I wrote fresh from a delightful roaring argument with my students about the issues at hand.

The study is a test of the tension-reduction view of alcohol consumption, conducted by Higgins and Marlatt (1973). Briefly, the subjects were made either highly anxious or not so anxious by the threat of electric shock, and were permitted access to alcohol as desired. If alcohol reduces tension and if people drink it because it does so (Cappell & Herman, 1972), then the anxious subjects should have drunk more. They did not.

Writing about this experiment, one of my better students gave it short shrift: "Surely not many al-

coholics are presented with such a threat under normal conditions."

Indeed. The threat of electric shock can hardly be "representative" of the dangers faced by anyone except electricians, hi-fi builders, and Psychology 101 students. What then? It depends! It depends on what kind of conclusion one draws and what one's purpose is in doing the study.

Higgins and Marlatt could have drawn this conclusion: "Threat of shock did not cause our subjects to drink in these circumstances. Therefore, it probably would not cause similar subjects to drink in similar circumstances either." A properly cautious conclusion, and manifestly trivial.

Or they could have drawn this conclusion: "Threat of shock did not cause our subjects to drink in these circumstances. Therefore, tension or anxiety probably does not cause people to drink in normal, real-world situations." That conclusion would be manifestly risky, not to say foolish; and it is that kind of conclusion which raises the issue of EV. Such a conclusion does assume that we can generalize from the simple and protected lab setting to the complex and dangerous real-life one and that the fear of shock can represent the general case of tension and anxiety. And let me admit again that we have been guilty of just this kind of foolishness on more than one occasion.

But that is not the conclusion Higgins and Marlatt drew. Their argument had an entirely different shape, one that changes everything. Paraphrased, it went thus: "Threat of shock did not cause our subjects to drink in these circumstances. Therefore, the tension-reduction hypothesis, which predicts that it should have done so, either is false or is in need of qualification." This is our old friend, the hypothetico-deductive method, in action. The important point to see is that the generalizability of the results, from lab to real life, is not claimed. It plays no part in the argument at all.

Of course, these findings may not require *much* modification of the tension-reduction hypothesis. It is possible—indeed it is highly likely—that there are tensions and tensions; and perhaps the nagging fears and self-doubts of the everyday have a quite different status from the acute fear of electric shock. Maybe alcohol does reduce these chronic fears and is taken, sometimes abusively, because it does so.[3] If these possibilities can be shown to be true, then we could sharpen the tension-reduction hypothesis, restricting it (as it is not restricted now) to certain kinds of tension and, perhaps, to certain settings. In short, we could advance our understanding. And the "artificial" laboratory findings would have contributed to that advance. Surely we cannot reasonably ask for more.

It seems to me that this kind of argument characterizes much of our research—much more of it than our critics recognize. In very many cases, we are not using what happens in the laboratory to "predict" the real world. Prediction goes the other way: Our theory specifies what subjects should do *in the laboratory.* Then we go to the laboratory to ask, Do they do it? And we modify our theory, or hang onto it for the time being, as results dictate. Thus we improve our theories, and—to say it again—it is these that generalize to the real world if anything does.

Let me turn to an example of another kind. To this point, it is artificiality of *setting* that has been the focus. Analogous considerations can arise, however, when one thinks through the implications of artificiality of, or bias in, the *sample.* Consider a case study.

A great deal of folklore, supported by some powerful psychological theories, would have it that children acquire speech of the forms approved by their culture—that is, grammatical speech—through the impact of parents' reactions to what they say. If a child emits a properly formed sentence (so the argument goes), the parent responds with approval or attention. If the utterance is ungrammatical, the parent corrects it or, at the least, withholds approval.

Direct observation of parent–child interactions, however, reveals that this need not happen. Brown and Hanlon (1970) report that parents react to the content of a child's speech, not to its form. If the sentence emitted is factually correct, it is likely to be approved by the parent; if false, disapproved. But whether the utterance embodies correct grammatical form has surprisingly little to do with the parent's reaction to it.

What kind of sample were Brown and Hanlon dealing with here? Families that (a) lived in Boston, (b) were well educated, and (c) were willing to have squadrons of psychologists camped in their living rooms, taping their conversations. It is virtually certain that the sample was biased even with respect to the already limited "population" of upper-class-Bostonian-parents-of-young-children.

Surely a sample like that is a poor basis from which to generalize to any interesting population. But what if we turn it around? We start with the theoretical proposition: Parents respond to the grammar of their children's utterances (as by making approval contingent or by correcting mistakes). Now we make the prediction: Therefore, the *parents*

[3] I should note, however, that there is considerable doubt about that as a statement of the general case. Like Harlow's experiment, the Higgins and Marlatt (1973) study articulates with a growing body of data from very different sources and settings, but all, in this case, calling the tension-reduction theory into question (cf. Mello & Mendelson, 1978).

we observe ought to do that. And the prediction is disconfirmed.

Going further, if we find that the children Brown and Hanlon studied went on to acquire Bostonian-approved syntax, as seems likely, then we can draw a further prediction and see it disconfirmed. If the theory is true, and if *these* parents do not react to grammaticality or its absence, then *these* children should not pick up grammatical speech. If they do so anyway, then parental approval is not necessary for the acquisition of grammar. And that is shown not by generalizing from sample to population, but by what happened *in the sample.*

It is of course legitimate to wonder whether the same contingencies would appear in Kansas City working-class families or in slum dwellers in the Argentine. Maybe parental approval/disapproval is a much more potent influence on children's speech in some cultures or subcultures than in others. Nevertheless, the fact would remain that the parental approval theory holds only in some instances and must be qualified appropriately. Again, that would be well worth knowing, and *this* sample of families would have played a part in establishing it.

The confusion here may reflect simple historical accident. Considerations of sampling from populations were brought to our attention largely by survey researchers, for whom the procedure of "generalizing to a population" is of vital concern. If we want to estimate the proportion of the electorate intending to vote for Candidate *X*, and if *Y*% of our sample intends to do so, then we want to be able to say something like this: "We can be 95% confident that *Y*% of the voters, plus or minus *Z*, intend to vote for *X*." Then the issue of representativeness is squarely before us, and the horror stories of biased sampling and wildly wrong predictions, from the *Literary Digest* poll on down, have every right to keep us awake at night.

But what has to be thought through, case by case, is whether that is the kind of conclusion we intend to draw. In the Brown and Hanlon (1970) case, nothing could be more unjustified than a statement of the kind, "We can be *W*% certain that *X*% of the utterances of Boston children, plus or minus *Y*, are true and are approved." The biased sample rules such a conclusion out of court at the outset. But it was never intended. The intended conclusion was not about a population but about a theory. That parental approval tracks content rather than form, in *these children,* means that the parental approval theory of grammar acquisition either is simply false or interacts in unsuspected ways with some attribute(s) of the home.

In yet other cases, the subjects are of interest precisely because of their unrepresentativeness. Washoe, Sarah, and our other special students are of interest because they are not representative of a language-using species. And with all the quarrels their accomplishments have given rise to, I have not seem them challenged as "unrepresentative chimps," except by students on examinations (I am not making that up). The achievements of mnemonists (which show us what *can* happen, rather than what typically *does*) are of interest because mnemonists are not representative of the rest of us. And when one comes across a mnemonist one studies that mnemonist, without much concern for his or her representativeness even as a mnemonist.

But what do students read? "Samples should always be as representative as possible of the population under study." "[A] major concern of the behavioral scientist is to ensure that the sample itself is a good representative [sic] of the population." (The sources of these quotations do not matter; they come from an accidental sample of books on my shelf.)

The trouble with these remarks is not that they are false—sometimes they are true—but that they are unqualified. Representativeness of sample is of vital importance for certain purposes, such as survey research. For other purposes it is a trivial issue.[4] Therefore, one must evaluate the sampling procedure in light of the purpose—separately, case by case.

Taking the Package Apart

Everyone knows that we make experimental settings artificial for a reason. We do it to control for extraneous variables and to permit separation of factors that do not come separately in Nature-as-you-find-it. But that leaves us wondering how, having stepped out of Nature, we get back in again. How do our findings apply to the real-life setting in all its complexity?

I think there are times when the answer has to be, "They don't." But we then may add, "Something else does. It is called understanding."

---

[4] There is another sense in which "generalizing to a population" attends most psychological research: One usually tests the significance of one's findings, and in doing so one speaks of sample values as estimates of population parameters. In this connection, though, the students are usually reassured that they can always define the population in terms of the sample and take it from there—which effectively leaves them wondering what all the flap was about in the first place.

Perhaps this is the place to note that some of the case studies I have presented may raise questions in the reader's mind that are not dealt with here. Some raise the problem of interpreting null conclusions; adequacy of controls for confounding variables may be worrisome; and the Brown and Hanlon (1970) study faced the problem of observer effects (adequately dealt with, I think; see Mook, 1982). Except perhaps for the last one, however, these issues are separate from the problem of external validity, which is the only concern here.

As an example, consider dark adaptation. Psychophysical experiments, conducted in restricted, simplified, ecologically invalid settings, have taught us these things among others:

1. Dark adaptation occurs in two phases. There is a rapid and rather small increase in sensitivity, followed by a delayed but greater increase.

2. The first of these phases reflects dark adaptation by the cones; the second, by the rods.

Hecht (1934) demonstrated the second of these conclusions by taking advantage of some facts about cones (themselves established in ecologically invalid photochemical and histological laboratories). Cones are densely packed near the fovea; and they are much less sensitive than the rods to the shorter visible wavelengths. Thus, Hecht was able to tease out the cone component of the dark-adaptation curve by making his stimuli small, restricting them to the center of the visual field, and turning them red.

Now let us contemplate the manifest ecological invalidity of this setting. We have a human subject in a dark room, staring at a place where a tiny red light may appear. Who on earth spends time doing that, in the world of affairs? And on each trial, the subject simply makes a "yes, I see it/no, I don't" response. Surely we have subjects who "behave as little like human beings as possible" (Bannister, 1966)—We might be calibrating a photocell for all the difference it would make.

How then do the findings apply to the real world? They do not. The task, variables, and setting have no real-world counterparts. What does apply, and in spades, is the understanding of how the visual system works that such experiments have given us. That is what we apply to the real-world setting—to flying planes at night, to the problem of reading X-ray prints on the spot, to effective treatment of night blindness produced by vitamin deficiency, and much besides.

Such experiments, I say, give us understanding of real-world phenomena. Why? Because the *processes* we dissect in the laboratory also operate in the real world. The dark-adaptation data are of interest because they show us a process that does occur in many real-world situations. Thus we could, it is true, look at the laboratory as a member of a class of "target" settings to which the results apply. But it certainly is not a "representative" member of that set. We might think of it as a limiting, or even *defining,* member of that set. To what settings do the results apply? The shortest answer is: to any setting in which it is relevant that (for instance) as the illumination dims, sensitivity to longer visible wavelengths drops out before sensitivity to short ones does. The findings do not represent a class of real-world phenomena; they define one.

Alternatively, one might use the lab not to explore a known phenomenon, but to determine whether such and such a phenomenon exists or can be made to occur. (Here again the emphasis is on what can happen, not what usually does.) Henshel (1980) has noted that some intriguing and important phenomena, such as biofeedback, could never have been discovered by sampling or mimicking natural settings. He points out, too, that if a desirable phenomenon occurs under laboratory conditions, one may seek to make natural settings mimic the laboratory rather than the other way around. Engineers are familiar with this approach. So, for instance, are many behavior therapists.

(I part company with Henshel's excellent discussion only when he writes, "The requirement of 'realism,' or a faithful mimicking of the outside world in the laboratory experiment, applies only to . . . hypothesis testing within the logico-deductive model of research" [p. 470]. For reasons given earlier, I do not think it need apply even there.)

The Drama of the Artificial

To this point, I have considered alternatives to the "analogue" model of research and have pointed out that we need not intend to generalize our results from sample to population, or from lab to life. There are cases in which we do want to do that, of course. Where we do, we meet another temptation: We may assume that in order to *generalize* to "real life," the laboratory setting should *resemble* the real-life one as much as possible. This assumption is the force behind the cry for "representative settings."

The assumption is false. There are cases in which the generalization from research setting to real-life settings is made all the stronger by the lack of resemblance between the two. Consider an example.

A research project that comes in for criticism along these lines is the well-known work on obedience by Milgram (1974). In his work, the difference between a laboratory and a real-life setting is brought sharply into focus. Soldiers in the jungles of Viet Nam, concentration camp guards on the fields of Eastern Europe—what resemblance do their environments bear to a sterile room with a shock generator and an intercom, presided over by a white-coated scientist? As a setting, Milgram's surely is a prototype of an "unnatural" one.

One possible reaction to that fact is to dismiss the work bag and baggage, as Argyle (1969) seems to do: "When a subject steps inside a psychological laboratory he steps out of culture, and all the normal rules and conventions are temporarily discarded and replaced by the single rule of laboratory culture—'do what the experimenter says, no matter how absurd or unethical it may be' " (p. 20). He goes on to cite Milgram's work as an example.

All of this—which is perfectly true—comes in a discussion of how "laboratory research can produce the wrong results" (Argyle, 1969, p. 19). The wrong results! But that is the whole point of the results. What Milgram has shown is how easily we can "step out of culture" in just the way Argyle describes—and how, once out of culture, we proceed to violate its "normal rules and conventions" in ways that are a revelation to us when they occur. Remember, by the way, that most of the people Milgram interviewed grossly underestimated the amount of compliance that would occur *in that laboratory setting.*

Another reaction, just as wrong but unfortunately even more tempting, is to start listing similarities and differences between the lab setting and the natural one. The temptation here is to get involved in count-'em mechanics: The more differences there are, the greater the external invalidity. Thus:

> One element lacking in Milgram's situation that typically obtains in similar naturalistic situations is that the experimenter had no real power to harm the subject if the subject failed to obey orders. The subject could always simply get up and walk out of the experiment, never to see the experimenter again. So when considering Milgram's results, it should be borne in mind that a powerful source of obedience in the real world was lacking in this situation. (Kantowitz & Roediger, 1978, pp. 387–388)

"Borne in mind" to what conclusion? Since the next sentence is "Nonetheless, Milgram's results are truly remarkable" (p. 388), we must suppose that the remarks were meant in criticism.

Now the lack of threat of punishment is, to be sure, a major difference between Milgram's lab and the jungle war or concentration camp setting. But what happened? An astonishing two thirds obeyed anyway. The force of the experimenter's authority was sufficient to induce normal decent adults to inflict pain on another human being, even though they could have refused without risk. Surely the absence of power to punish, though a distinct difference between Milgram's setting and the others, only adds to the drama of what he saw.

There are other threats to the external validity of Milgram's findings, and some of them must be taken more seriously. There is the possibility that the orders he gave were "legitimized by the laboratory setting" (Orne & Evans, 1965, p. 199). Perhaps his subjects said in effect, "This is a scientific experiment run by a responsible investigator, so maybe the whole business isn't as dangerous as it looks." This possibility (which is quite distinct from the last one, though the checklist approach often confuses the two) does leave us with nagging doubts about the generalizability of Milgram's findings. Camp guards and jungle fighters do not have this cognitive escape hatch available to them. If Milgram's subjects did say "It must not be dangerous," then his conclusion—people are surprisingly willing to inflict danger under orders—is in fact weakened.

The important thing to see is that the checklist approach will not serve us. Here we have two differences between lab and life—the absence of punishment and the possibility of discounting the danger of obedience. The latter difference weakens the impact of Milgram's findings; the former strengthens it. Obviously we must move beyond a simple count of differences and think through what the effect of each one is likely to be.

Validity of What?

Ultimately, what makes research findings of interest is that they help us understand everyday life. That understanding, however, comes from theory or the analysis of mechanism; it is not a matter of "generalizing" the findings themselves. This kind of validity applies (if it does) to statements like "The hunger-reduction interpretation of infant attachment will not do," or "Theory-driven inferences may bias first impressions," or "The Purkinje shift occurs because rod vision has *these* characteristics and cone vision has *those.*" The validity of these generalizations is tested by their success at prediction and has nothing to do with the naturalness, representativeness, or even nonreactivity of the investigations on which they rest.

Of course there are also those cases in which one does want to predict real-life behavior directly from research findings. Survey research, and most experiments in applied settings such as factory or classroom, have that end in view. Predicting real-life behavior is a perfectly legitimate and honorable way to use research. When we engage in it, we do confront the problem of EV, and Babbie's (1975) comment about the artificiality of experiments has force.

What I have argued here is that Babbie's comment has force *only* then. If this is so, then external validity, far from being "obviously our ideal" (Campbell & Stanley, 1967), is a concept that applies only to a rather limited subset of the research we do.

A Checklist of Decisions

I am afraid that there is no alternative to thinking through, case by case, (a) what conclusion we want to draw and (b) whether the specifics of our sample or setting will prevent us from drawing it. Of course there are seldom any fixed rules about how to "think through" anything interesting. But here is a sample of questions one might ask in deciding whether the usual criteria of external validity should even be considered:

As to the sample: Am I (or is he or she whose work I am evaluating) trying to estimate from sam-

ple characteristics the characteristics of some population? Or am I trying to draw conclusions not about a population, but about a theory that specifies what *these* subjects ought to do? Or (as in linguistic apes) would it be important if *any* subject does, or can be made to do, this or that?

As to the setting: Is it my intention to predict what would happen in a real-life setting or "target" class of such settings? Our "thinking through" divides depending on the answer.

The answer may be no. Once again, we may be testing a prediction rather than making one; our theory may specify what ought to happen in *this* setting. Then the question is whether the setting gives the theory a fair hearing, and the external-validity question vanishes altogether.

Or the answer may be yes. Then we must ask, Is it therefore necessary that the setting be "representative" of the class of target settings? Is it enough that it be *a* member of that class, if it captures processes that must operate in all such settings? If the latter, perhaps it should be a "limiting case" of the settings in which the processes operate—the simplest possible one, as a psychophysics lab is intended to be. In that case, the stripped-down setting may actually *define* the class of target settings to which the findings apply, as in the dark-adaptation story. The question is only whether the setting actually preserves the processes of interest,[5] and again the issue of external validity disappears.

We may push our thinking through a step further. Suppose there are distinct differences between the research setting and the real-life target ones. We should remember to ask: So what? Will they weaken or restrict our conclusions? Or might they actually strengthen and extend them (as does the absence of power to punish in Milgram's experiments)?

Thinking through is of course another warm, fuzzy phrase, I quite agree. But I mean it to contrast with the cold creepies with which my students assault research findings: knee-jerk reactions to "artificiality"; finger-jerk pointing to "biased samples" and "unnatural settings"; and now, tongue-jerk imprecations about "external invalidity." People are already far too eager to dismiss what we have learned (even that biased sample who come to college and elect our courses!). If they do so, let it be for the right reasons.

[5] Of course, whether an artificial setting does preserve the process can be a very real question. Much controversy centers on such questions as whether the operant-conditioning chamber really captures the processes that operate in, say, the marketplace. If resolution of that issue comes, however, it will depend on whether the one setting permits successful predictions about the other. It will not come from pointing to the "unnaturalness" of the one and the "naturalness" of the other. There is no dispute about that.

REFERENCES

Argyle, M. *Social interaction.* Chicago: Atherton Press, 1969.

Babbie, E. R. *The practice of social research.* Belmont, Calif.: Wadsworth, 1975.

Bannister, D. Psychology as an exercise in paradox. *Bulletin of the British Psychological Society,* 1966, *19,* 21–26.

Bickman, L. Social roles and uniforms: Clothes make the person. *Psychology Today,* July 1974, pp. 49–51.

Brown, R., & Hanlon, C. Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley, 1970.

Brunswik, E. Representative design and probabilistic theory in a functional psychology. *Psychological Review,* 1955, *62,* 193–217.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1967.

Cappell, H., & Herman, C. P. Alcohol and tension reduction: A review. *Quarterly Journal of Studies on Alcohol,* 1972, *33,* 33–64.

Deese, J. *Psychology as science and art.* New York: Harcourt Brace Jovanovich, 1972.

Hecht, S. Vision II: The nature of the photoreceptor process. In C. Murchison (Ed.), *Handbook of general experimental psychology.* Worcester, Mass.: Clark University Press, 1934.

Henshel, R. L. The purposes of laboratory experimentation and the virtues of deliberate artificiality. *Journal of Experimental Social Psychology,* 1980, *16,* 466–478.

Higgins, R. L., & Marlatt, G. A. Effects of anxiety arousal on the consumption of alcohol by alcoholics and social drinkers. *Journal of Consulting and Clinical Psychology,* 1973, *41,* 426–433.

Kantowitz, B. H., & Roediger, H. L., III. *Experimental psychology.* Chicago: Rand McNally, 1978.

Mello, N. K., & Mendelson, J. H. Alcohol and human behavior. In L. L. Iverson, S. D. Iverson, & S. H. Snyder (Eds.), *Handbook of psychopharmacology: Vol. 12. Drugs of abuse.* New York: Plenum Press, 1978.

Milgram, S. *Obedience to authority.* New York: Harper & Row, 1974.

Mook, D. G. *Psychological research: Strategy and tactics.* New York: Harper & Row, 1982.

Nisbett, R. E., & Ross, L. *Human inference: Strategies and shortcomings in social judgment.* New York: Century, 1980.

Orne, M. T., & Evans, T. J. Social control in the psychological experiment: Anti-social behavior and hypnosis. *Journal of Personality and Social Psychology,* 1965, *1,* 189–200.