

*Confusion about the meaning of validity in quasi-experimental research can be addressed by carefully relabeling types of validity. Internal validity can more aptly be termed "local molar causal validity." More tentatively, the "principle of proximal similarity" can be substituted for the concept of external validity.*

## *Relabeling Internal and External Validity for Applied Social Scientists*

*Donald T. Campbell*

On the one hand, there have been widespread expressions of dissatisfaction with the distinction between internal validity and external validity and suggestions for its revision or elimination. In Cook and Campbell (1979), Tom Cook began a review of this literature, and he now has in preparation a much more systematic and extensive one (Cook, 1985; Cook and Campbell, in preparation). This chapter avoids that task.

On the other hand, both those who have enthusiastically adopted the distinction and those who oppose it, have most frequently redefined it to epitomize all the differences between pure laboratory experimentation and field tryouts of ameliorative programs. Are you, dear reader, perhaps one who has done so? Half of my own students fail to answer the following question correctly: When one adds a placebo control group in a pharmaceutical experiment, is this done to improve internal or external validity? As I read Campbell (1957) and Campbell and Stanley (1966), the correct answer is external validity.

---

This chapter is based in large part on the final report prepared under contract number SSN 552-12-4531 with the U.S. Department of Health and Human Services. The opinions expressed are those of the author and are not to be taken as government policy.

The meanings compiled in dictionaries are properly based upon usage, and the meanings ascribed to specialist terms are based upon usage in the relevant specialty. On these grounds, the term *internal validity* now means similarity to the pure treatment (rule-of-one-variable), fully controlled, laboratory experiment. Since that is not what we had in mind, we need to try again, with new terms.

It will help to remind ourselves of the dialectical motivation that led to introduction of the distinction between internal and external validity. In the 1950s, the training in research methods that social psychologists received was dominated by Fisherian analysis of variance statistics, as though random assignment to treatment were the only methodological control that needed to be taught. (How the physiological psychologists kept alive an earlier model is a matter for another discussion.) In dialogue with this lopsided and complacent emphasis, we wished to point out that in out-of-doors social experimentation, there were a lot of threats to validity that randomization did not take care of and that the teaching of research design should be expanded to cover these other threats, which we classified as issues of external validity. It was against the overwhelming dominance of Fisher's randomized assignment models and an implicit, complacent assumption that meticulous care in this regard took care of all experimental validity problems that we were reacting. Thus, threats to external validity came to be defined as threats controlled for by random assignment to treatment. And, backhandedly, threats to internal validity were, initially and implicitly, those for which random assignment did control. This overlaps with the laboratory ideal, to be sure, but it excludes purity of the treatment variable. This was most nearly made clear by the inclusion of generalization to other treatments in external validity from the very beginning.

With my approval, Tom Cook has fuzzed up this simple distinction and added four threats to internal validity which randomization does not control (Cook and Campbell, 1979). It is symptomatic of problems with the concept that we now believe that at least one of these threats, resentful demoralization of respondents receiving less desirable treatments, if not all four, might well be grouped with construct validity of treatments. In the hypothetical case that Cook (Cook and Campbell, 1979) used to illustrate the threat, the posttest difference between experimental and control groups was a product of no change in the experimental group and entirely due to the resentful demoralization of the control group. In terms of the internal validity concept that I used in scoring the placebo-control-group item that I give to students, one could answer affirmatively the internal validity question, Did the experimental contrast as a total package in its specific setting cause a real difference? and relegate to construct validity (external validity in Campbell and Stanley, 1966) the question of how to interpret that validity noted difference.

In June 1984, Tom Cook and I spent three days planning a revision or new book. We tentatively agreed that *internal validity* needed relabeling. We

went round and round on alternative labels. We are not satisfied with our present one, *local molar causal validity*, and it will probably not survive this chapter. The choice of terms is again due to the present stage of the dialogue. In continentalese, it has “historicist dialectical indexicality” (Campbell, 1982, p. 327). That is, the choice of terms is a product of the particular argument of this historical moment. *Molar* is used in specific reaction to the connotation of the theoretically “pure,” “simple” experimental treatment. *Local* is used in reaction to a generalizability concept that may be a holdover from the logical positivist’s so-called covering law model. Why the current rejection of positivism (in which, via Campbell and Fiske, 1959, I was a pioneer) should include covering laws I do not fully understand. Covering laws would seem like such nice things to have from any point of view on science. Perhaps it is that they implicitly require at the beginning of an inquiry knowledge that at best will only be available at completion (see also d’Andrade, 1986).

### **Local Molar (Pragmatic, Atheoretical) Causal Validity**

For the applied scientist, local molar causal validity is a first crucial issue and the starting point for other validity questions. For example, did this complex treatment package make a real difference in this unique application at this particular place and time? By *molar* we connote recognition that the treatment is often a very complex hodgepodge (from the point of view of abstract analytic theoretical science), which has been put together by expert clinical judgment, not on the basis of the already proven efficacy of its theoretically pure components (main effects plus interactions). By *molar* we also connote an interest in evaluating this complex treatment as it stands, rather than first testing its hundred or so components one at a time, or in a hundred-variables-by-several-levels-each randomized ANOVA experiment. The molar approach assumes that clinical practice, participant observation, and epidemiological studies already have accumulated some wisdom, suggesting treatments that are worth further testing as molar packages. If these packages turn out to have striking molar efficacy, we will, of course, be interested in further studies, both clinically and theoretically guided, that will help us to determine which of several conjectured major components is most responsible for the effect. These later studies in turn will still be using complex packages, rather than testing theoretically pure variables in isolation or in experimentally controlled higher-order interaction. Pure-variable science can, of course, be a source of treatment packages (as in brain metabolite therapy for children diagnosed as potentially schizophrenic due to metabolite abnormality), but in preventive intervention these, too, will inevitably be a part of a complex social system of diagnosis and delivery. They will need to be tested as intervention packages under conditions of eventual application, or under facsimiles of such situations, that have been chosen both for clarity of scientific inference and for similarity to target conditions of application.

By *local* we indicate the strategy illustrated in pilot testing: Let's see if it really works in some one setting and time. If it does, later on we can explore the boundaries of its efficacy in other locales and with specialized populations. If it does not, we may be appropriately discouraged from further trials, even though it might conceivably work in some other setting.

While the molar local causal validity of the applied social scientist may be a far cry from the agenda of basic science, most of the problems of validity—and thus the methodology for the establishment of validity—are shared. Thus (if the applied research problem is not thereby abandoned), the two traditions of experimental control—experimental isolation and randomized assignment to treatments—are also ideal ways of establishing local molar causal validity. For applied ameliorative research under the field conditions of application, random assignment to treatments usually is the optimal approach to local molar causal validity, although not for external and construct validities.

*Molar* and *local* could both be taken as implying no generalization at all, conceptualizing a validly demonstrated cause-effect relationship that we do not as yet know how to generalize. The causal relationship would be known locally and molarly, but there would be no validated theory of it that would guide generalization to other interventions, measures, populations, settings, or times. This is, of course, an exaggeration. The theories and hunches used by those who put the therapeutic package together must, of course, be regarded as corroborated, however tentatively, if there is an effect of local, molar validity in the expected direction. Nonetheless, this exaggeration may serve to remind us that very frequently in physical science (and probably in social science as well) causal puzzles (that is, effects that are dependable but not understood) are the driving motor for new and productive theorizing. We must back up from the current overemphasis on theory first.

Basic scientists put a premium on clarity of causal inference and hence limit, trim, and change problems so that they can be solved with scientific precision given the current state of the art. Other causal hypotheses are postponed until the state of the art and theory development make them precisely testable. This strategy is not available to applied scientists. They should stay with the mandated problem, doing the best they can to achieve scientific validity but (in order to stay with the problem) often making use of methods providing less precision of causal inference where necessary. Thus, we applied social scientists need not only randomized experiments and strong quasi-experiments but also case studies, ethnography, participant observation, gossip collection from informants, hermeneutics, and so forth. Ideally, these materials will be used to provide the context necessary for valid estimation of the seriousness of the threats to validity and for valid interpretation of the results of formal experimentation, but if need be they may be used alone. We need these not because the social sciences seek a different kind of validity than other sciences do, but rather because to stay with our problems we must

use techniques that, while improving the validity of our research, nonetheless provide less clarity of causal inference than would a retreat to narrowly specified variables under laboratory control. While using these techniques of the humanities, staying in real-world, nonlaboratory settings, the critical tools of threats to validity and plausible rival hypotheses are still central (Becker, 1979; Cook and Reichardt, 1979).

### Relabeling External Validity

Even farther from stability or even transient consensus is a complementary reconceptualization of external validity. The following discussion, however dogmatic it may be, is a tentative and incomplete tryout of the principle of proximal similarity for that role. While that obviously will not quite do as a heading under which the present threats to construct validity and external validity fall logically, it may help to shake us free of past conceptualizations, as a way station en route to something more satisfactory. Here I move to this concept by stages.

*Generalizing from the Local Without Representative Sampling.* Akin to the relabeling of *internal validity* as *local molar causal validity* is a reformulation of the concept of external validity or generalizability set forth by Campbell and Stanley (1966). This has, of course, already been done by Cook (Cook and Campbell, 1979), who explicitly separated out and reconceptualized the issues of generalizing to other nonidentical treatments, now called *construct validity of causes*, and of generalizing from the outcome measures employed to other measures of effects, now called *construct validity of effects*. Remaining in the Cook and Campbell (1979) residual category of external validity is the validity with which one can generalize to other persons, settings, and times. Such generalizations should also be made on the basis of theory and thus they, too, should be reconceptualized as construct validities. That is, the validity of generalizations to other persons, settings, and future (or past) times would be a function of the validity of the theory involved, plus the accuracy of the theory-relevant knowledge of the persons, settings, and future periods to which one wanted to generalize (for example, to which one wanted to apply an intervention with demonstrated local molar causal validity).

This perspective has already moved us far from the widespread concept that one can solve generalizability problems by representative sampling from a universe specified in advance. Such an approach is obviously impossible for sampling from future occasions. However, the statistical technology and practical possibility is available for persons and for specified setting units, such as schools, schoolrooms, factories, hospitals, retail stores, cities, and counties. While national samples along these lines are often called for in evaluations mandated by Congress, it turns out that nearly all the high-quality scientific program studies (such as guaranteed annual income, housing allowances, school vouchers, coverage of psychotherapy by health

insurance, and so forth) have chosen illustrative samples, exemplifying the target population in informal judgmental ways, employing samples of feasibility if not samples of convenience. For the New Jersey negative income tax experiment (Kershaw and Fair, 1976; Watts and Rees, 1977), the researchers gave the idea of a nationwide randomly selected sample of low-income households very careful consideration before deciding upon a few areas in New Jersey and a portion of Pennsylvania selected for both feasibility and theoretical reasons of an unquantified, general, and informal sort. I strongly endorse this approach. I believe it characterizes physical science as well as the most valid and useful of applied science.

This is, to be sure, an unpracticed ideal. But, it is so out of keeping with what we know of science that it should be removed even from our philosophy of science. A consideration of the time dimension will help to show that it is utterly unreasonable. In the physical sciences, the presumption that there are no interactions with time (except those of daily, lunar, seasonal, and other cycles) has proved to be a reasonable one. But, for the social sciences, a consideration of the characteristics of potentially relevant populations shows that changes over time (for example, a thirty-year comparison of college students) produce differences fully as large as synchronous social class and subcultural differences. To sample representatively from our intended universe of generalization would require representative sampling in time, an obvious impossibility.

More typical of science is the case of Nicholson and Carlisle. Taking in May 1800 a very parochial and idiochronic sample of Soho water, inserting into it a very biased sample of copper wire, into which flowed a very local electrical current, they obtained hydrogen at one electrode and oxygen at the other and uninhibitedly generalized to all the water in the world for all eternity. It was a hypothetical generalization, to be sure, rather than a proven fact. There have been by now many studies of the effect of impurities in the water upon hydrolysis, but these studies, too, have been based on very biased samples. The idea of a representative sampling of all the waters of the world, or even of all the waters of England, never occurred even as an idea. The very concept of impurities, of distinguishing the contents of water as "pure" stuff and alien materials, is one that would never have emerged had a representative sampling approach to water been employed. In the successful sciences, generalizations have never been inductive in the sense of summarizing what has been observed within the bounds of the generalization, but instead they have always been presumptive, albeit guided by prior laws. The limitations on generalization have emerged from efforts to check on an initial bold generalization in nonrepresentative ways. Scientists assumed that hydrolysis held true universally until it was shown otherwise.

In this light, had we achieved one, there would be no need to apologize for a successful psychology of college sophomores, or even of Northwestern University coeds, or of Wistar strain white rats. Exciting and

powerful laws would then be presumed to hold for all humans or all vertebrates at all times until specific applications of that presumption proved wrong. We already are at this latter stage, but even here a representative sampling of species or school populations is not the answer. Theory-guided, dimensional explorations, as in comparing primates that vary widely in evolutionary development, are in the typical path of science (Campbell, 1969).

In program evaluation, I at least, recommend formal abandonment of the goal of nationally representative sample selection. Once there are interventions of such well-established effectiveness that the decision is made to adopt them nationally, sample census data on population distribution and sample censuses on schools, hospitals, or other distribution facilities to be employed and on labor and space costs can be employed for budgetary planning purposes. I do not anticipate that cross-validation of an intervention's effectiveness on a strictly representative national sample would ever be cost-beneficial or needed. If it were employed, it would be for administrative reasons, not for applied scientific validity.

*The Principle of Proximal Similarity.* The first presentation of external validity (Campbell, 1957) was entirely in terms of generalization to other treatments, measures, populations, settings, and times. As stated earlier, I feel we need something more appropriate than the generalization rhetoric and the solution of it by representative sampling from a universe designated in advance. This rhetoric is greatly reduced in Cook and Campbell (1979) through Tom Cook's notion of construct validity of causes and effects. In this shift, the validity of theoretical interpretation replaces atheoretical generalization to other treatments and measures. A shift has been made from a positivist phenomenalism to a fallibilist realism in which all treatments and measures are regarded as imperfect proxy variables for latent causes and effects.

But, even in this chapter, the rhetoric of "generalizing to" still persists, both in what has gone before and in what will follow. Somehow, I feel we need to preserve the valid aspects of our problem statement in a conceptual framework still more emancipated, still more characteristic of the coherence strategy of belief revision (Quine, 1951; Campbell, 1966, 1978), which we employ *faute de mieux*, even if we cling to a correspondence goal and meaning for the concept of truth. But, this chapter does not achieve that goal.

Under the principle of proximal similarity I would like to provide a metatheoretical basis for justifying a seemingly atheoretical rationale and approach to the generalization of findings. I do this ambivalently, because one of the attractive summaries of our new contrast is to regard local molar causal validity as atheoretical and construct or external validity as theoretical. Perhaps the principle of proximal similarity merely describes the route to theory-based generalization, given the multiattribute contexts with which we

must begin and from which we can only be released by degrees of experimental isolation and control that are inaccessible in social settings.

While I believe that the principle of proximal similarity applies to pure science also, I want to make an argument for it specific to applied social science. In so doing I borrow from some earlier papers (Campbell, 1972; Campbell, 1973; Raser and others 1970). It was Harrod's (1956) effort to justify induction that moved me to this conceptualization. While I judge that he failed, as all such efforts must fail, his work introduces a profoundly different understanding of the presuppositions underlying scientists' efforts at inductive inference. For the earlier postulate that nature is orderly, Harrod substitutes the presupposition that nature is "sticky," "viscous," proximally autocorrelated in space, time, and probably  $n$ -dimensional attribute space, with adjacent points more similar (as a rule) than nonadjacent ones.

The most important practical justification of the principle (and of the need for confirming in practice the efficacy of social ameliorations) comes from the fact that our experience in generalizing social science findings shows that higher-order interactions abound, precluding unqualified generalization of our principles not only from laboratory to laboratory but especially from laboratory to field application.

It is most convenient to explain this in terms of the model of analysis of variance. Consider multiple dimensions of experimental variation  $A$ ,  $B$ ,  $C$ ,  $D$ , and so forth, each of which occurs in several degrees of strength, with (in the simplest design) each combination of strengths being employed. (Thus, if there were four dimensions,  $A$ ,  $B$ ,  $C$ , and  $D$ , each of which had three strengths, there would be eighty-one different treatment packages. In addition to these treatment or independent variables, there is at least one dependent variable in terms of which the results of the treatments are measured. Let us call this  $X$ . For our present purposes, two major types of outcome need to be distinguished: main effects and interactions. If a main effect for  $A$  is found on  $X$ , then we have what could be called a *ceteris paribus* law:  $B$ ,  $C$ ,  $D$  being held constant at any level, the same rule relating  $A$  to  $X$  is found: For example, the more  $A$ , the more  $X$ . Where interactions are found, the relations are complexly contingent. For example, in an  $A$ - $B$  interaction, there may be a separate rule relating  $A$  and  $X$  for each different level of  $B$  (for example, if  $B$  is high, the more  $A$  the more  $X$ , but if  $B$  is low, the more  $A$  the less  $X$ ). Much more complex (higher-order) interactions can also occur, such as an  $A$ - $B$ - $C$  interaction in which the  $A$ -to- $X$  rule is different for each combination of  $B$  and  $C$ .

Interactions, where they occur in the absence of main effects, represent highly limited and qualified generalizations. It is typical of the history of the physical sciences that many strong main effects have been found—generalizations conceivably true independent of time and place and the status of other variables. While eventually, in fine detail, the laws were found to be more complex, there was nonetheless a rich experience of discovering



approximate laws of nature that could be stated without specifying the conditions on the infinitude of other potentially relevant variables.

There is no compelling evidence so far that the social sciences are similarly situated. If we take the one social science that uses the analysis of variance approach, experimental social psychology, the general finding is of abundant higher-order interactions and rare main effects. Even where we get main effects, it is certainly often due to the failure to include the additional dimensions (*E*, *F*, *G*, and so forth) that would have produced interactions. Frequently we are unable to replicate findings from one university laboratory to another, indicating an interaction with some unspecified difference in the laboratory settings or in the participants.

If such multiple factorial experiments can be regarded as experiments in generalization, they give us grounds for great caution, particularly when we generalize the expectation that, had we included dimensions *E*, *F*, *G*, and *H* or *Y* and *Z* in our experiment, the *A*–*X* relationship might well have shown interactions with some or all of them, too. The high rate of interactions on the variables that we have explored must make us expect something similar for the many variables that we have not explored.

Any given experiment can be regarded as holding constant at one particular level every one of the innumerable variables on which no experimental variation is introduced, each of which is like a single level of a potential experiment in which two or more levels of the same variable were systematically employed. We can guess with confidence that the farther apart the two values of *B* (or *E* or *Z*), the more likely it is that *B* will interact with the *A*–*X* relationship. (An empirical exploration of this might well be worth making. Data from complex experiments using three or more levels of a given treatment could be reanalyzed as two-level experiments, some as wide-range, using the two extreme levels and disregarding the intermediate, others as narrow-range, using adjacent levels from the original experiment.)

In anticipation of the outcome of such studies and in common with the intuition of most scientists, let us assume as a general rule that the larger the range of values on the background variable, the more likely these variables are to have strong interactions with the *A*–*X* relationship under study. Or, to put it more simply, as scientists we generalize with most confidence to applications most similar to the setting of the original research. When generalizing from our laboratory-based theory to a real-world social-ameliorative program, the values on all dimensions differ widely, and new interaction effects, as yet unexplored, become extremely likely.

Intuitively, we already use this principle of proximal similarity in many ways, and we can self-consciously use it in more. When it comes to disseminating a new ameliorative program of local molar causal validity, we will apply it with most confidence where treatment, setting, population, desired outcome, and year are closest in some overall way to the original

program treatment. In contrast, for research on the limits of generalization, exploratory contrasts should be sought out for cross-validation that differ as much as possible from the first intervention in population, setting, and so forth while remaining within the legislatively targeted populations and problems. Purposive sampling for maximum exploration of generalizability on conceptualized dimensions will be substituted for a population-representative sampling.

In the new contrast, external and construct validities involve theory. Local molar causal validity does not. While this contrast is weakened in the principle of proximal similarity, I still want to retain it. The principle of proximal similarity is normally (and it should be) implemented on the basis of expert intuition. The use of the term *construct* in the expression *construct validity of causes and effects* (Cook and Campbell, 1979) may too strongly connote formal theory. Nevertheless, most philosophers or at least most logicians may well agree with Nelson Goodman (1972) that any concept of overall similarity is meaningless or incoherent, since there are potentially an infinite number of attribute dimensions on which such similarity could be computed. Our intuitive expectations about what dimensions are relevant are theory-like, even if they are not formally theoretical. Moreover, clinical experience, prior experimental results, and formal theory are very appropriate guides for efforts to make the exploration of the bounds of generalizability more systematic.

### Nonconclusion

The material presented in this chapter is self-consciously inconclusive. It is a dialectical reaction, or overreaction. Let us hope that the overall iteration is headed for convergence.

### References

- Becker, H. S. "Do Photographers Tell the Truth?" In T. D. Cook and C. S. Reichardt (eds.), *Qualitative and Quantitative Methods in Evaluation Research*. Vol. 1. Beverly Hills, Calif.: Sage, 1979.
- Campbell, D. T. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin*, 1957, 54, 297-312.
- Campbell, D. T. "Pattern Matching as an Essential in Distal Knowing." In K. R. Hammond (ed.), *The Psychology of Egon Brunswik*. New York: Holt, Rinehart and Winston, 1966.
- Campbell, D. T. "Prospective: Artifact and Control." In R. Rosenthal and R. Rosnow (eds.), *Artifact in Behavior Research*. New York: Academic Press, 1969.
- Campbell, D. T. "Herskovits, Cultural Relativism, and Metascience." In M. J. Herskovits, *Cultural Relativism*. New York: Random House, 1972.
- Campbell, D. T. "The Social Scientist as Methodological Servant of the Experimenting Society." *Policy Studies Journal*, 1973, 2, 72-75.
- Campbell, D. T. "Qualitative Knowing in Action Research." In M. Brenner, P. Marsh, and M. Brenner (eds.), *The Social Contexts of Method*. London: Croon Helm, 1978.

- Campbell D. T. "Experiments as Arguments." *Knowledge: Creation, Diffusion, Utilization*, 1982, 3, 327-337.
- Campbell, D. T. and Fiske, D. W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin*, 1959, 56, 81-105.
- Campbell, D. T., and Stanley, J. C. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966.
- Cook, T. D. "Recent Attacks on Well-Known Validity Distinctions: An Appreciative Rejoinder." Paper presented at the annual convention of the Midwestern Psychological Association, Chicago, May 3, 1985.
- Cook, T. D., and Campbell, D. T. *Quasi-Experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin, 1979.
- Cook, T. D., and Campbell, D. T. "Quasi-Experimental Research: Conceptual and Design Issues After a Quarter Century of Practice and Criticism," forthcoming.
- Cook T. D., and Reichardt, C. S. *Qualitative and Quantitative Methods in Evaluation Research*. Beverly Hills, Calif.: Sage, 1979.
- D'Andrade, R. "Three Scientific World Views and the Covering Law Model." In D. W. Fiske and R. A. Shweder (eds.), *Metatheory in Social Science*. Chicago: University of Chicago Press, 1986.
- Goodman, N. "Likeness." In N. Goodman (ed.), *Problems and Projects*. Indianapolis, Ind: Bobbs-Merrill, 1972.
- Harrod, R. F. *Foundations of Inductive Logic*. London: Macmillan, 1956.
- Kershaw, D., and Fair, J. *The New Jersey Income Maintenance Experiment*, Vol. 1: *Operations, Surveys, and Administration*, New York: Academic Press, 1976.
- Quine, W. V. "Two Dogmas of Empiricism." *Philosophical Review*, 1951, 60, 20-43.
- Raser, J. R., Campbell, D. T., and Chadwick, R. W. "Gaming and Simulation for Developing Theory Relevant to International Relations." *General Systems Research*. Vol. 15. Ann Arbor, Mich.: Society for General Systems Research, 1970.
- Watts, H. W., and Rees, A. (eds.). *The New Jersey Income Maintenance Experiment*. Vol. 2: *Labor-Supply Responses*. Vol. 3: *Expenditures, Health, and Social Behavior and the Quality of the Evidence*. New York: Academic Press, 1977.

*Donald T. Campbell is past president of the American Psychological Association; member of the National Academy of Sciences; professor of social relations, psychology, and education at Lehigh University; and recipient in 1977 of the Myrdal Prize in Science of the Evaluation Research Society.*