QUASI-EXPERIMENTAL DESIGN[1]

Donald T. Campbell

Northwestern University

This phrase refers to the application of an experimental mode of analysis and interpretation to bodies of data not meeting the full requirements of experimental control because experimental units are not assigned at random to at least two "treatment" conditions. The settings to which it is appropriate are those of experimentation in social settings, including planned interventions such as specific communications, persuasive efforts, changes in conditions and policies, efforts at social remediation, etc. Unplanned conditions and events may also be analyzed in this way where an exogeneous variable has such discreteness and abruptness as to make appropriate its consideration as an experimental treatment applied at a specific point in time to a specific population. When properly done, when attention is given to the specific implications of the specific weaknesses of the design in question, quasi-experimental analysis can provide a valuable extension of the experimental method.

While efforts to interpret field data as experiments go back much farther, the first prominent methodology of this kind in the social sciences was Chapin's Ex Post Facto Experiment (Chapin & Queen, 1937; Chapin, 1955; Greenwood, 1945), although it should be noted that due to the failure to control regression artifacts, this mode of analysis is no longer regarded as acceptable. The American Soldier volumes (Stouffer et al., 1949) provide prominent analyses of the effects of specific military experiences, where it is implausible that differences in selection explain the results. Thorndike's efforts to demonstrate the effects of specific course work upon other intellectual achievements provide an excellent early model (e.g., Thorndike & Woodworth, 1901; Thorndike & Ruger, 1923). Extensive analysis and review of this literature are provided elsewhere (Campbell, 1957; 1963; Campbell & Stanley, 1963) and serve as the basis for the present abbreviated presentation.

The core requirement of a "true" experiment lies in the experimenter's ability to apply at least two experimental treatments in complete independence of the prior states of the materials (persons, etc.) under study. This independence makes resulting differences interpretable as effects of the differences in treatment. In the social sciences this independence of prior status is assured by randomization in assignments to treatments. Experiments meeting these requirements, and thus representing "true" experiments, are much more possible in the social sciences than is generally realized. Wherever, for example, the treatments can be applied to individuals or small units (such as precincts or classrooms) without the respondents' being aware of experimentation or that other units are getting different treatments, very elegant experimental control can be achieved. An increased acceptance by administrators of randomization as the democratic method of allocating scarce resources (be these new housing, therapy, or fellowships) will make possible field experimentation in many settings. Where innovations are to be introduced throughout a social system, and where the introduction cannot in any event be simultaneous, a use of randomization in the staging can provide an experimental comparison of the new and the old, using the groups receiving the delayed introduction as controls. Nothing in this article should be interpreted as minimizing the importance of increasing the use of true experimentation. However, where true experimental design with random assignment of persons to treatments is not possible, due to ethical considerations, lack of power, or in feasibility, application of quasi-experimental analysis has much to offer.

The social sciences must do the best they can with the possibilities open to them. Inferences must frequently be made from data lacking complete control. Too often a scientist trained in experimental method rejects out of hand any research in which complete control is lacking. Yet in practice no experiment is perfectly executed, and the practicing scientist overlooks those imperfections which seem to him to offer no plausible rival explanation of the results. In the light of modern philosophies of science, no experiment ever proves a theory, it merely probes it. Seeming proof results from that condition in which there is no available plausible rival hypothesis to explain the data. The general program of quasi-experimental analysis is to specify and examine those plausible rival explanations of the results which are provided by the uncontrolled variables. A failure of control which does not in fact provide a plausible rival interpretation is not regarded as invalidating.

It is well to remember that we do make assured causal inferences in many settings not involving randomization: (The earthquake caused the brick building to crumble; the automobile crashing into it caused the telephone pole to break; the language patterns of the older models

---

[1]The preparation of this review was supported in part by Project C-998, Contract 3-20-001, with the Media Research Branch, Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act. This symposium presentation is essentially the same as the current draft of my article for the International Encyclopedia of the Social Sciences.

and mentors caused this child to speak English rather than Kwakiutl; etc.) While these are all potentially erroneous inferences, they are of the same type as experimental inferences. We are confident that were we to intrude experimentally, we could confirm the causal laws involved. Yet they have been made assuredly by a nonexperimenting observer. This assurance is due to the effective absence of other plausible causes. Consider the inference as to crashing auto and the telephone pole: we rule out combinations of termites and wind because the other implications of these theories (e.g., termite tunnels and debris in the wood, wind records at nearby weather stations) do not occur. Spontaneous splintering of the pole by happenstance coincident with the auto's onset does not impress us as a rival, nor would it explain the damage to the car, etc. Analogously in quasi-experimental analysis, tentative causal interpretation of data may be made where the interpretation in question squares with the data and where other rival interpretations have been rendered implausible.

For the evaluation of data series as quasi-experiments, a set of twelve frequent threats to validity have been developed. These may be regarded as the important classes of frequently plausible rival hypotheses which good research design seeks to rule out. All will be presented briefly even though not all are employed in the evaluation of the designs used illustratively here.

Fundamental to this listing is a distinction between internal validity and external validity. Internal validity is the basic minimum without which any experiment is uninterpretable: did in fact the experimental treatments make a difference in this specific experimental instance? External validity asks the question of generalizability: to what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds, in that features increasing one may jeopardize the other. While internal validity is the sine qua non, and while the question of external validity, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal.

Relevant to internal validity are eight different classes of extraneous variables which, if not controlled in the experimental design, might produce effects mistaken for the effect of the experimental treatment. These are: 1. History: the other specific events occurring between a first and second measurement in addition to the experimental variable. 2. Maturation: processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing tireder, and the like. 3. Testing: the effects of taking a test upon the scores of a second testing. 4. Instrumentation: in which changes in the calibration of a measuring instrument or changes in the observers or scorers used may produce changes in the obtained measurements. 5. Statistical regression: operating where groups have been selected on the basis of their extreme scores. 6. Selection: biases resulting in differential recruitment of respondents for the comparison groups. 7. Experimental mortality: the differential loss of respondents from the comparison groups. 8. Selection-maturation interaction: In certain of the multiple-group quasi-experimental designs, such as the nonequivalent control group design, such interaction is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Factors jeopardizing external validity or representativeness are: 9. The reactive or interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected. 10. Interaction effects between selection bias and the experimental variable. 11. Reactive effects of experimental arrangements, which would preclude generalization about the effect of the experimental variable for persons being exposed to it in nonexperimental settings. 12. Multiple-treatment inference, a problem wherever multiple treatments are applied to the same respondents, and a particular problem for one-group designs involving equivalent time-samples or equivalent materials samples.

Perhaps the simplest quasi-experimental design is the One-Group Pretest-Posttest Design, $\underline{O}_1 \underline{X} \underline{O}_2$ (where O represents measurement or observation, and X represents the experimental treatment). This common design patently leaves uncontrolled the internal validity threats of History, Maturation, Testing, Instrumentation, and, if selected as extreme on $\underline{O}_1$, Regression. There may be situations in which the analyst could decide that none of these represented plausible rival hypotheses in his setting: A log of other possible change-agents might provide no plausible ones, the measurement in question might be nonreactive (Campbell, 1957), the time span too short for maturation, too spaced for fatigue, etc. However, the sources of invalidity are so numerous that a more powerful quasi-experimental design would be preferred. Several of these can be constructed by adding features to this simple one. The Interrupted Time-Series Experiment utilizes a series of measurements providing multiple pretests and posttests, e.g.: $\underline{O}_1 \underline{O}_2 \underline{O}_3 \underline{O}_4 \underline{X} \underline{O}_5 \underline{O}_6 \underline{O}_7 \underline{O}_8$. If in this series, $\underline{O}_4 - \underline{O}_5$ shows a rise greater than found elsewhere, then Maturation, Testing, and Regression are no longer plausible, in that they would predict equal or greater rises for $O_1 - O_2$, etc. Instrumentation may well be controlled too, although in institutional settings a change of administration policy is often accompanied by a change in record-keeping standards. Observers and participants may be focused on the occurrence of X, and may fail to take into consideration

changes in rating standards, etc. History remains the major threat, although in many settings it would not offer a plausible rival interpretation. If one had available a parallel time series from a group not receiving the experimental treatment, but exposed to the same extraneous sources of influence, and if this control time series failed to show the exceptional jump from $\underline{O}_4$ to $\underline{O}_5$, then the plausibility of History as a rival interpretation would be greatly reduced. We may call this the Multiple Time-Series Design.

Another way of improving the One-Group Pretest-Posttest Design is to add a "Nonequivalent Control Group." (Were the control group to be randomly assigned from the same population as the experimental group, we would, of course, have a true, not quasi, experimental design.) Depending on the similarities of setting and attributes, if the nonequivalent control group fails to show a gain manifest in the experimental group, then History, Maturation, Testing, and Instrumentation are controlled. In this popular design, the frequent effort to "correct" for the lack of perfect equivalence by matching on pretest scores is absolutely wrong (e.g., Thorndike, 1942; Hovland et al., 1949; Campbell & Clayton, 1961), as it introduces a regression artifact. Instead, one should live with any initial pretest differences, using analysis of covariance, or graphic presentation. Remaining uncontrolled is the Selection-Maturation Interaction, i.e., the possibility that the experimental group differed from the control group not only in initial level, but also in its autonomous maturation rate. In experiments on psychotherapy and on the effects of specific coursework this is a very serious rival. Note that it can be rendered implausible by use of a time series of pretests for both groups, thus moving again to the Multiple Time-Series Design.

There is not space here to present adequately even these four quasi-experimental designs, but perhaps the strategy of adding specific observations and analyses to check on specific threats to validity has been illustrated. This is carried to an extreme in the Recurrent Institutional Cycle Design (Campbell & McCormack, 1957; Campbell & Stanley, 1963), in which longitudinal and cross-sectional measurements are combined with still other analyses to assess the impact of indoctrination procedures, etc., through exploiting the fact that essentially similar treatments are being given to new entrants year after year or cycle after cycle. Other quasi-experimental designs covered in Campbell & Stanley (1963) include two more single-group designs (the Equivalent Time-Samples Design and the Equivalent Materials Design), Counterbalanced or Rotational Designs, Separate Sample Pretest-Posttest Designs, Regression-Discontinuity Analysis, the Panel Impact Design (see also Campbell & Clayton, 1961), and the Cross-Lagged Panel Correlation, which is related to Lazarsfeld's Sixteen-Fold Table (see especially Campbell, 1963).

Related to the program of quasi-experimental analysis are those efforts to achieve causal inference from correlational data. Note that while correlation does not prove causation, most causal hypotheses imply specific correlations, and thus examination of these probes, tests, or edits the causal hypothesis. Further, as Simon and Blalock have emphasized (e.g., Blalock, 1961), certain causal models specify uneven patterns of correlation. Thus the $A \rightarrow B \rightarrow C$ model implies that r_{AC} be smaller than r_{AB} or r_{BC}. However, the use of partial correlations or the use of Wright's (1920) path analysis are rejected by the present writer as tests of the model because of the requirement that the "cause" be totally represented in the "effect." In the social sciences it will never be plausible that the "cause" has been measured without unique error and that it also totally lacks unique systematic variance not shared with the "effect." More appropriate would be Lawley's (1940) test of the hypothesis of single-factoredness. Only if single-factoredness can be rejected would the causal model as represented by its predicted uneven correlations pattern be the preferred interpretation.

A word needs to be said about tests of significance for quasi-experimental designs. There has come from several competent social scientists the argument that since randomization has not been used, tests of significance assuming randomization are not relevant. The attitude of the present writer is on the whole in disagreement. However, some aspects of the protest are endorsed: Good experimental design is needed for any comparison inferring change, whether or not tests of significance are used, even if only photographs, graphs, or essays are being compared. In this sense, experimental design is independent of tests of significance. More importantly, tests of significance have come to be taken as thoroughgoing proof. In vulgar social science usage, finding a "significant difference" is apt to be taken as proving the author's basis for predicting the difference, forgetting the many other plausible rival hypotheses explaining a significant difference which quasi-experimental designs leave uncontrolled. Certainly the valuation of tests of significance in some quarters needs demoting. Further, the use of tests of significance designed for the evaluation of a single comparison becomes much too lenient when dozens, hundreds, or thousands of comparisons have been sifted, and this is still common usage. And in a similar manner, the author's decision as to which of his studies is publishable, and the editor's decision as to which of the manuscripts is acceptable, further biases the sampling basis. In all of these ways, reform is needed.

However, when a quasi-experimenter has compared the results from two intact classrooms employed in a sampling of convenience, sample size, small-sample instability, a chance difference, is certainly one of the many plausible rival hypotheses which must be considered, even if only one. If each class had but five students we would interpret the fact that 20% more in the experimental class showed increases in favorable-

ness with much less interest than if each class had 500 students. In this case there is available an elaborate formal theory for the plausible rival hypothesis of chance fluctuation. This theory involves assumptions of randomness, which are quite appropriately present when we reject the null model of random association in favor of a hypothesis of systematic difference between the two classes. If we find a "significant difference," the test of significance will not, of course, tell us whether the two classes differed because one saw the experimental movie, or for some selection reason associated with class topic, time of day, etc., which might have interacted with rate of autonomous change, pre-test instigated changes, reactions to commonly experienced events, etc. But such a test of significance will help us rule out this 13th plausible rival hypothesis, that there is no difference here at all that a model of purely chance assignment could not account for as a vagary of sampling. Note that our statement of probability level is in this light a statement of the plausibility of this rival hypothesis, which always has some plausibility, however faint. In this orientation, a practice of stating the probability in descriptive detail seems preferable to using but a single apriori decision criterion.

REFERENCES

Blalock, H.M. 1964 Causal inferences in non-experimental research. Chapel Hill: The University of North Carolina Press, 1964.

Campbell, D.T. 1957 Factors relevant to validity of experiments in social settings. Psychological Bulletin 54:297-312.

Campbell, D.T. 1963 From description to experimentation: Interpreting trends as quasi-experiments. Pages 212-242 in Harris, C.W. (editor), Problems in measuring change. Madison, Wis.: University of Wisconsin Press.

Campbell, D.T.; and Clayton, K.N. 1961 Avoiding regression effects in panel studies of communication impact. Studies in Public Communication No. 3, 99-118.

Campbell, D.T.; and McCormack, Thelma H. 1957 Military experience and attitudes toward authority. American Journal of Sociology 62:482-490.

Campbell, D.T.; and Stanley, J.C. 1963 Experimental and quasi-experimental designs for research on teaching. Pages 171-246 in Gage, N.L. (editor), Handbook of research on teaching. Chicago: Rand McNally.

Chapin, F.S. 1955 Experimental designs in sociological research. New York: Harper. (Rev. ed.)

Chapin, F.S.; and Queen, S.A. 1937 Research memorandum on social work in the depression. New York: Social Science Research Council, Bulletin 39, 1937.

Greenwood, E. 1945 Experimental sociology: A study in method. New York: King's Crown Press.

Hovland, C.I.; Lumsdaine, A.A.; and Sheffield, F.C. 1949 Experiments on mass communication. Princeton, N.J.: Princeton University Press.

Lawley, C.N. 1940 The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh 60:64-82.

Stouffer, S.S. (editor) 1949 The American Soldier. Princeton, N.J.: Princeton University Press. Vols. I and II.

Thorndike, R.L. 1942 Regression fallacies in the matched groups experiment. Psychometrika 7:85-102.

Thorndike, E.L.; and Ruger, G.J. 1923 The effect of first-year Latin upon knowledge of English words of Latin derivation. School and Society 81:260-270, 417-418.

Thorndike, E.L.; and Woodworth, R.S. 1901 The influence of improvement in one mental function upon the efficiency of other functions. Psychological Review 8:247-261, 384-395, 553-564.

Wright, S. 1920 Correlation and causation. Journal of Agricultural Research 20:557-585.