NERA

THE RESEARCHER

ANNUAL JOURNAL OF THE
NORTHEASTERN EDUCATIONAL RESEARCH
ASSOCIATION

NOREEN B. GARMAN, EDITOR

SPECIAL EDITION:
New Scholars and Their Work

September, 1982
Volume 1, Number 1

Introduction

For the past two years the Board of Directors of the Northeastern Educational Research Association has envisioned the publication of a special edition of *The Researcher* as an annual document. They decided that a journal format would support the purpose of the organization as stated in the Constitution, that is, to promote educational research by the provision of a forum and by the interchange of information through publications. In addition, the leadership of NERA has expressed a responsibility to help new scholars in the field with their research efforts. This strong commitment is represented here in the 1982 first special edition. Primary authors are new scholars; and, in some cases, senior colleagues have co-authored with them. The printing format for the journal was determined by financial exigencies, (a postulate of the times). In the coming years, as the journal becomes a NERA tradition, decisions by the members about a conventional focus and format will reflect the high quality of this fine association.

Major credit for the spirit of this endeavor must be given to President, Janice Gibson, whose intellect and organizational skill gave shape to the 1982 publications. With her leadership, it has been a privilege to serve as *The RESEARCHER* editor for the past year.

Noreen Garman, Editor
University of Pittsburgh

Designing Designs for Research

William M. K. Trochim
Cornell University
Douglas A. Land
Cornell University

Much contemporary social research is devoted to examining whether a program, treatment, or manipulation causes some outcome or result. For example, we might wish to know whether a new educational program causes subsequent achievement score gains, whether a special work-release program for prisoners causes lower recidivism rates, whether a novel drug causes a reduction in symptoms, and so on. Cook and Campbell (1979) argue that three conditions must be met before we can infer that such a cause-effect relation exists:

1. Covariation. Changes in the presumed cause must be related to changes in the presumed effect. Thus, if we introduce, remove, or change the level of a treatment or program, we should observe some change in the outcome measures.
2. Temporal Precedence. The presumed cause must occur prior to the presumed effect.
3. No Plausible Alternative Explanations. The presumed cause must be the only reasonable explanation for changes in the outcome measures. If there are other factors which could be responsible for changes in the outcome measures we cannot be confident that the presumed cause-effect relationship is correct.

In most social research the third condition is the most difficult to meet. Any number of factors other than the treatment or program could cause changes in outcome measures. Campbell and Stanley (1966) and later, Cook and Campbell (1979) list a number of common plausible alternative explanations (or, threats to internal validity). For example, it may be that some historical event which occurs at the same time that the program or treatment is instituted was responsible for the change in the outcome measures; or, changes in record keeping or measurement systems which occur at the same time as the program might be falsely attributed to the program. The reader is referred to standard research methods texts for more detailed discussions of threats to validity.

This paper is primarily heuristic in purpose. Standard social science methodology textbooks (Cook and Campbell

1979; Judd and Kenny, 1981) typically present an array of research designs and the alternative explanations which these designs rule out or minimize. This tends to foster a "cookbook" approach to research design - an emphasis on the selection of an available design rather than on the construction of an appropriate research strategy. While standard designs may sometimes fit real-life situations, it will often be necessary to "tailor" a research design to minimize specific threats to validity. Furthermore, even if standard textbook designs are used, an understanding of the logic of design construction in general will improve the comprehension of these standard approaches. This paper takes a structural approach to research design. While this is by no means the only strategy for constructing research designs, it helps to clarify some of the basic principles of design logic.

Minimizing Threats to Validity

Good research designs minimize the plausible alternative explanations for the hypothesized cause-effect relationship. But such explanations may be ruled out or minimized in a number of ways other than by design. The discussion which follows outlines five ways to minimize threats to validity, one of which is by research design:

1. By Argument. The most straightforward way to rule out a potential threat to validity is to simply argue that the threat in question is not a reasonable one. Such an argument may be made either a priori or a posteriori, although the former will usually be more convincing than the latter. For example, depending on the situation, one might argue that an instrumentation threat is not likely because the same test is used for pre and post test measurements and did not involve observers who might improve, or other such factors. In most cases, ruling out a potential threat to validity by argument alone will be weaker than the other approaches listed below. As a result, the most plausible threats in a study should not, except in unusual cases, be ruled out by argument only.
2. By Measurement or Observation. In some cases it will be possible to rule out a threat by measuring it and demonstrating that either it does not occur at all or occurs so minimally as to not be a strong alternative explanation for the cause-effect relationship. Consider for example a study of the effects of an advertising campaign on subsequent sales of a particular product. In such a study, history (i.e., the occurrence of other events which might lead to an increased desire to purchase the product) would be a plausible alternative explanation. For example, a change in the local economy, the removal of a competing product from the market, or similar events could cause an increase in product sales. One might attempt to minimize such threats by measuring local economic indicators and the availability and sales of competing products. If there is no change in these measures coincident with the onset of the advertising campaign, these threats would be considerably minimized. Similarly, if one is studying the effects of special mathematics training on math achievement scores of children, it might be useful to observe everyday classroom behavior in order to verify that students were not receiving any additional math training to that provided in the study.
3. By Design. Here, the major emphasis is on ruling out alternative explanations by adding treatment or control groups, waves of measurement, and the like. This topic will be discussed in more detail below.
4. By Analysis. There are a number of ways to rule out alternative explanations using statistical analysis. One interesting example is provided by Jurs and Glass (1971). They suggest that one could study the plausibility of an attrition or mortality threat by conducting a two-way analysis of variance. One factor in this study would be the original treatment group designations (i.e., program vs. comparison group), while the other factor would be attrition (i.e., dropout vs. non-dropout group). The dependent measure could be the pretest or other available pre-program measures. A main effect on the attrition factor would be indicative of a threat to external validity or generalizability, while an interaction between group and attrition factors would point to a possible threat to internal validity. Where both effects occur, it is reasonable to infer that there is a threat to both internal and external validity.

   The plausibility of alternative explanations might also be minimized using covariance analysis. For example, in a study of the effects of "workfare" programs on social welfare case loads, one plausible alternative explanation might be the status of local economic conditions. Here, it might be possible to construct a measure of economic conditions and include that measure as a covariate in the statistical analysis. One must be careful when using covariate adjustments of this type-"perfect" covariates do not exist in most social research and the use of imperfect covariates will not completely adjust for potential alternative explanations. Nevertheless causal assertions are likely to be strengthened by demonstrating that treatment effects occur even after adjusting on a number of good covariates.
5. By Preventive Action. When potential threats are anticipated they can often be ruled out by some type of preventive action. For example, if the program is a desireable one, it is likely that the comparison group would feel jealous or demoralized. Several actions can be taken to minimize the effects of these attitudes including offering the program to the comparison group upon completion of the study or using program and comparison groups which have little opportunity for contact and communication. In addition, auditing methods and quality control can be used to track potential experimental dropouts or to insure the standardization of measurement.

The five categories listed above should not be considered mutually exclusive. The inclusion of measurements designed to minimize threats to validity will obviously be related to the design structure and is likely to be a factor in the analysis. A good research plan should, where possible, make use of multiple methods for reducing threats. In general, reducing a particular threat by design or preven-

tive action will probably be stronger than by using one of the other three approaches. The choice of which strategy to use for any particular threat is complex and depends at least on the cost of the strategy and on the potential seriousness of the threat.

Design Construction

Basic Design Elements. Most research designs can be constructed from four basic elements:

1. Time. A causal relationship, by its very nature, implies that some time has elapsed between the occurrence of the cause and the consequent effect. While for some phenomena the elapsed time might be measured in microseconds and therefore might be unnoticeable to a casual observer, we normally assume that the cause and effect in social science arenas do not occur simultaneously. In design notation we indicate this temporal element horizontally - whatever symbol is used to indicate the presumed cause would be placed to the left of the symbol indicating measurement of the effect. Thus, as we read from left to right in design notation we are reading across time. Complex designs might involve a lengthy sequence of observations and programs or treatments across time.
2. Program(s) or Treatment(s). The presumed cause may be a program or treatment under the explicit control of the researcher or the occurrence of some natural event or program not explicitly controlled. In design notation we usually depict a presumed cause with the symbol "X". When multiple programs or treatments are being studied using the same design, we can keep the programs distinct by using subscripts such as "X^1" or "X^2". For a comparison group (i.e., one which does not receive the program under study) no "X" is used.
3. Observation(s) or Measure(s). Measurements are typically depicted in design notation with the symbol "O". If the same measurement or observation is taken at every point in time in a design, then this "O" will be sufficient. Similarly, if the same set of measures is given at every point in time in this study, the "O" can be used to depict the entire set of measures. However, if different measures are given at different times it is useful to subscript the "O" to indicate which measurement is being given at which point in time.
4. Groups or Individuals. The final design element consists of the intact groups or the individuals who participate in various conditions. Typically, there will be one or more program and comparison groups. In design notation, each group is indicated on a separate line. Furthermore, the manner in which groups are assigned to the conditions can be indicated by an appropriate symbol at the beginning of each line. In this paper "R" will represent a group which was randomly assigned, "N" will depict a group which was nonrandomly assigned (i.e., a nonequivalent group or cohort) and a "C" will indicate that the group was assigned using a cutoff score on a measurement.

Perhaps the easiest way to understand how these four basic elements become integrated into a design structure is to give several examples. One of the most commonly used designs in social research is the two-group pre-post design which can be depicted as:

N O X O
N O O

There are two lines in the design indicating that the study was comprised of two groups. The two groups were nonrandomly assigned as indicated by the "N". Both groups were measured before the program or treatment occurred as indicated by the first "O" in each line. Following this pre-observation, the group in the first line received a program or treatment while the group in the second line did not. Finally, both groups were measured subsequent to the program. Another common design is the posttest-only randomized experiment. The design can be depicted as:

R X O
R O

Here, two groups are randomly selected with one group receiving the program and one acting as a comparison. Both groups are measured after the program is administered.

Expanding a Design. We can combine the four basic design elements in a number of ways in order to arrive at a specific design which is appropriate for the setting at hand. One strategy for doing so begins with the basic causal-relationship:

X O

This is the most simple design in causal research and serves as a starting point for the development of better strategies. When we add to this basic design we are essentially expanding one of the four basic elements described above. Each possible expansion has implications both for the cost of the study and for the threats which might be ruled out.

I. Expanding Across Time. We can add to the basic design by including additional observations either before or after the program or, by adding or removing the program or different programs. For example, we might add one or more pre-program measurements and achieve the following design:

O O X O

The addition of such pretests provides a "baseline" which, for instance, helps to assess the potential of a maturation or testing threat. If a change occurs between the first and second pre-program measures, it is reasonable to expect that similar change might be seen between the second pretest and the posttest even in the absence of the program. However, if no change occurs between the two pretests, one might be more confident in assuming that maturation or testing is not a likely alternative explanation for the cause-effect relationship which was hypothesized. Similarly, additional post-program measures could be added. This would be useful for determining whether an immediate program effect decays over time, or whether there is a lag in time between the initiation of the program and the occurence of an effect. We might also add and remove the program over time:

O X O O X O

This is one form of the ABAB design which is frequently used in clinical psychology and psychiatry. The design is particularly strong against a history threat. When the program

is repeated it is less likely that unique historical events can be responsible for replicated outcome patterns.

II. Expanding Across Programs. We have just seen that we can expand the program by adding it or removing it across time. Another way to expand the program would be to partition it into different levels of treatment. For example, in a study of the effect of a novel drug on subsequent behavior, we might use more than one doseage of the drug:

O X1 O
O X2 O

This design is an example of a simple factorial design with one factor having two levels. Notice that group assignment is not specified indicating that any type of assignment might have been used. This is a common strategy in a "sensitivity" or "parametric" study where the primary focus is one th effects obtained at various program levels. In a similar manner, one might expand the program by varying specific components of it across groups. This might be useful if one wishes to study different modes of the delivery of the program, different sets of program materials and the like. Finally, we can expand the program by using theoretically polarized or "opposite" treatments. A comparison group can be considered one example of such a polarization. Another might involve use of a second program which is expected to have an opposite effect on the outcome measures. A strategy of this sort provides evidence that the outcome measure is sensitive enough to differentiate between different programs.

III. Expanding Across Observations. At any point in time in a research design it is usually desirable to collect multiple measurements. For example we might add a number of similar measures in order to determine whether the results of these converge. Or, we might wish to add measurements which theoretically should not be affected by the program in question in order to demonstrate that the program discriminates between effects. Strategies of this type for achieving convergent and discriminant validity of measures are discussed in Campbell and Fiske (1959). Another way to expand the observations is by proxy measurements. Assume that we wish to study a new educational program but neglected to take pre-program measurements. We might use a standardized achievement test for the posttest and grade point average records as a proxy measure of student achievement prior to the initiation of the program. Finally, we might also expand the observations through the use of "recollected" measures. Again, if we were conducting a study and had neglected to administer a pretest or desired information in addition to the pretest information, we might ask participants to recall how they felt or behaved prior to the study and use this information as an additional measure. Different measurement approaches obviously yield data of different quality. What is advocated here is the use of multiple measurements rather than reliance on only a single strategy.

IV. Expanding Across Groups. Often, it will be to our advantage to add additional groups to a design in order to rule out specific threats to validity. For example, consider the following pre-post two-group randomized experimental design:

R O X O
R O O

If this design were implemented within a single institution where members of the two groups were in contact with each other one might expect that intergroup communication, group rivalry, or demoralization of a group which gets denied a desirable treatment or gains an undesirable one might pose threats to the validity of the causal inference. In such a case, one might add an additional nonequivalent group from a similar institution which consists of persons unaware of the original two groups:

R O X O
R O O
N O O

In a similar manner, whenever nonequivalent groups are used in a study it will usually be advantageous to have multiple replications of each group. The use of many nonequivalent groups helps to minimize the potential of a particular selection bias affecting the results. In some cases it may be desirable to include the norm group as an additional group in the design. Norming group averages are available for most standardized achievement tests for example, and might comprise an additional nonequivalent control group. Cohort groups might also be used in a number of ways. For example, one might use a single measure of a cohort group to help rule out a testing threat:

R O X O
R O O
N O

In this design, the randomized groups might be sixth graders from the same school year while the cohort might be the entire sixth grade from the previous academic year. This cohort group did not take the pretest and, if they are similar to the randomly selected control group, would provide evidence for or against the notion that taking the pretest had an effect on posttest scores. We might also use pre-post cohort groups:

N O X O
N O O
N O O

Here, the treatment group consists of sixth graders, the first comparison group of seventh graders in the same year, and the second comparison group consists of the following year's sixth graders (i.e., the fifth graders during the study year). Strategies of this sort are particularly useful in nonequivalent designs where selection bias is a potential problem and where routinely-collected institutional data is available. Finally, one other approach for expanding the groups involves partitioning groups with different assignment strategies. For example, one might randomly divide nonequivalent groups, or select nonequivalent subgroups from randomly assigned groups. An example of this sort involving the combination of random assignment and assignment by a cutoff is discussed in detail below.

A Simple Strategy for Design Construction. Considering the basic elements of a research design or the possibilities

for expansion are not alone sufficient. We need to be able to integrate these elements with an overall strategy. Furthermore, we need to decide which potential threats are best handled by design rather than by argument, measurement, analysis, or preventive action.

While no definitive approach for designing designs exists, we might suggest a tentative strategy based on the notion of expansion discussed above. First, we begin the designing task by setting forth a design which depicts the simple hypothesized causal relationship. Second, we deliberately *overexpand* this basic design by expanding across time, program, observations, and groups. At this step, the emphasis is on accounting for as many likely alternative explanations as possible using the design. Finally, we then scale back this overexpanded version considering the effect of eliminating each design component. It is at this point that we face the difficult decisions concerning the costs of each design component and the advantages of ruling out specific threats using other approaches.

There are several advantages which result from using this type of approach to design construction. First, we are forced to be explicit about the decisions which are made. Second, the approach is "conservative" in nature. The strategy minimizes the chance of our overlooking a major threat to validity in constructing our design. Third, we arrive at a design which is "tailored" to the situation at hand. Finally, the strategy is cost-efficient. Threats which can be accounted for by some other, less costly, approach need not be accounted for in the design itself.

An Example of a Hybrid Design

Some of the ideas discussed above can be illustrated in an example. The design in question is drawn from an earlier discussion by Boruch (1975). To our knowledge, this design has never been used, although it has strong features to commend it.

Let us assume that we wish to study the effects of a new compensatory education program on subsequent student achievement. The program is designed to help students who are poor in reading to improve in those skills. We can begin then with the simple hypothesized cause-effect relationship:

X O

Here, the "X" represents the reading program and the "O" stands for a reading achievement test. We decide that it is desirable to add a pre-program measure so that we might investigate whether the program "improves" reading test scores. We also decide to expand across groups by adding a comparison group. At this point we have the typical:

O X O
O O

The next problem concerns how the two groups will be assigned. Since the program is specifically designed to help students who need special assistance in reading, we rule out random assignment because it would require denying the program to students in need. We had considered the possibility of offering the program to one randomly assigned group in the first year and to the control group in the second, but ruled that out on the grounds that it would require two years of program expenses and the denial of a potentially helpful program for half of the students for a period of a year. Instead we decide to assign students by means of a cutoff score on the pretest. All students scoring below a preselected percentile on the reading pretest would be given the program while those above that percentile would act as controls (i.e., the regression-discontinuity design). However, previous experience with this strategy (Trochim, in press) has shown us that it is difficult to adhere to a single cutoff score for assignment to group. We are especially concerned that teachers or administrators will allow students who score slightly above the cutoff point into the program because they have little confidence in the ability of the achievement test to make fine distinctions in reading skills for children who score very close to the cutoff. To deal with this potential problem, we decide to partition the groups using a particular combination of assignment by a cutoff and random assignment:

C O X O
R O X O
R O O
C O O

In this design we have set up two cutoff points. All those scoring below a certain percentile are assigned to the treatment group automatically by this cutoff. All those scoring above another higher percentile are automatically assigned to the comparison group by this cutoff. Finally, all those who fall in the interval between the cutoffs on the pretest are randomly assigned to either the treatment or comparison groups.

There are several advantages to this strategy. It directly addresses the concern to teachers and administrators that the test may not be able to discriminate well between students who score immediately above or below a cutoff point. For example, a student whose true ability in reading would place him near the cutoff might have a bad day and therefore might be placed into the treatment or comparison group by chance factors. The design outlined above is defensible. We can agree with the teachers and administrators that the test is fallible. Nevertheless, since we need some criterion to assign students to the program, we can argue that the fairest approach would be to assign borderline cases by lottery. In addition, by combining two excellent strategies (i.e., the randomized experiment and the regression-discontinuity) we can analyze results separately for each and address the possibility that design factors might bias results.

There are many other worthwhile considerations not mentioned in the above scenario. For example, instead of using simple randomized assignment within the cutoff interval, we might use a weighted random assignment so that students scoring lower in the interval have a greater probability of being assigned to the program. In addition, we might consider expanding the design in a number of other ways, by including double pretests or multiple posttests; multiple measures of reading skills; additional replications of the program or variations of the program; and additional groups such as norming groups, controls from other schools, and the like. Nevertheless, this brief example serves to illustrate the advantages of explicitly constructing a research design to meet the specific needs of a particular situation.

The Nature of Good Design

Throughout the design construction task it is important to have in mind some endpoint, some criteria which we should try to achieve before finally accepting a design strategy. The criteria discussed below are only meant to be suggestive of the characteristics found in good research design. It is worth noting that all of these criteria point to the need to individually tailor research designs rather than accepting standard textbook strategies as is.

1. Theory-Grounded. Good research strategies reflect the theories which are being investigated. Where specific theoretical expectations can be hypothesized these are incorporated into the design. For example, where theory predicts a specific treatment effect on one measure but not on another, the inclusion of both in the design improves discriminant validity and demonstrates the predictive power of the theory.
2. Situational. Good research designs reflect the settings of the investigation. This was illustrated above where a particular need of teachers and administrators was explicitly addressed in the design strategy. Similarly, intergroup rivalry, demoralization, and competition might be assessed through use of additional comparison groups who are not in direct contact with the original groups.
3. Feasible. Good designs can be implemented. The sequence and timing of events are carefully thought out. Potential problems in measurement, adherence to assignment, database construction and the like, are anticipated. Where needed, additional groups or measurements are included in the design to explicitly correct for such problems.
4. Redundant. Good research designs have some flexibility built into them. Often, this flexibility results from duplicaton of essential design features. For example, multiple replications of a treatment help to insure that failure to implement the treatment in one setting will not invalidate the entire study.
5. Efficient. Good designs strike a balance between redundancy and the tendency to overdesign. Where it is reasonable, other, less costly, strategies for ruling out potential threats to validity are utilized.

This is by no means an exhaustive list of the criteria by which we can judge good research design. Nevertheless, goals of this sort help to guide the researcher toward a final design choice and emphasize important components which should be included.

The development of a theory of research methodology for the social sciences has largely occurred over the past half century and most intensively within the past two decades. It is not surprising, in such a relatively recent effort, that an emphasis on a few standard research designs has occurred. Nevertheless, by moving away from the notion of "design selection" and towards an emphasis on design construction, there is much to be gained in our understanding of design principles and in the quality of our research.

References

Boruch, R. F. Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods and Research,* 4:1, 1975, 31-53.

Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 56, 1959, 81-105.

Campbell, D. T. and Stanley, J. C. *Experimental and Quasi-Experimental Designs for Research,* Chicago: Rand McNally, 1966.

Cook, T. D. and Campbell, D. T. *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally, 1979. Judd, C. M. and Kenny, D. A. *Estimating the Effects of Social Interventions.* Cambridge: Cambridge University Press, 1981.

Jurs, S. G. and Glass, G. V. The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *The Journal of Experimental Education,* 40:1, 1971, 62-66.

Trochim, W. Different designs, similar programs, different results: Methodologically-based discrepancies in the results of compensatory education evaluations. *Evaluation Review,* in press.

Evaluating Thinking

Hope Hartman-Haas

Newark Board of Education

Although thinking is perhaps the most basic of all cognitive skills, it has received scant attention from educators and educational researchers. Educators may neglect thinking, at least in part, because it has not been accorded legitimacy as an important intellectual skill which would benefit from direct instruction. Effective thinking, apparently, is assumed to emerge spontaneously during the course of development. Even if an educator were to decide that direct instruction in thinking is desirable, there is a scarcity of curricular materials specifically designed to improve thinking. However, even if such materials were more readily available, how would educators evaluate the effectiveness of their instruction? Both educators and educational researchers are constrained by the current paucity of measurement devices and techniques for effectively evaluating thinking. Those that exist reflect the gap between psychological knowledge and educational practice.

Traditional Approaches to Evaluating Thinking

In psychology and education, thinking has often been equated with intelligence. During the early part of the twentieth century, around the time of the first World War, intelligence was one of the most popular topics of academic concern. The mental testing movement was born during that