Experimental Design in Educational Research

DAVID R. KRATHWOHL

LET US PROJECT an ambitious plan and try to follow it as far as time permits. First of all, let us see if we can obtain some perspective on the research process so as to see what it is that researchers attempt. Second, let us determine the function of experimental design in that process. Third, let us list some of the variables which can be taken into account by a good experimental design. Finally, let us look at a control group type of research design and see how it takes into account the variables discussed above. Obviously, there is so much to include that we shall not be able to cover all of this material in detail.

What is it that we are trying to do in research? The terms "experimentation" and "research" mean many things to different people. To some they mean trying something out to see how well they like it. To scientists they mean careful work and precise methodology. "Research in education" too often takes on the connotation of the former rather than the latter, but there is a considerable difference between merely trying something out and observing its effect, and the careful measurement and analysis we except in research. When one tries something out on an informal basis, one more or less unconsciously evaluates it against what was used in the situation before. This informal evaluation represents a chain of reasoning to determine the better method. In research, one consciously establishes a basis for comparison and delineates the basis on which the comparison is to be made. Basically, one tries to build a very tight chain of argument to the effect that something may be true. This is the first aspect of the perspective on the research process.

A second aspect of the perspective is gained when we ask, "Of what does this chain of argument consist?" It usually starts with a hunch that some relation exists (for instance, that decentralization

David R. Krathwohl is Director of the Bureau of Educational Research Services, College of Education, Michigan State University, East Lansing, Michigan.

of the library results in its greater use) or that some practice is true (a particular way of cataloging leads to greater accessibility). At this step, we are formulating an hypothesis around which we hope to build a chain of reasoning which will show whether or not the hypothesis is true. As the next step, we must gather some observations which would permit us to ascertain the truth or falsity of our hunch or hypothesis. As we decide where we will make our observations, we make decisions about the sample which we shall use. When we make the decisions about what we shall observe, we define operationally the terms in our hypothesis, definitions which are the measures of the variables in our study. We choose in what setting we shall observe the phenomenon in question, making sure that we observe the correct thing and that what we observe is not affected by some extraneous variable that is not part of our hypothesis. At this point, we are developing our experimental design. Thus in many instances our experimental design contrasts observations in an experimental setting with those taken in a like setting where the experiment was not carried on. (This is the contrast between an experimental and a control group.) Finally, we must have some way of evaluating our observations to see whether what we expected did indeed occur and that this occurrence was not a chance happening, that is, a happening which might have occurred because of the particular sample chosen. A statistical model assists us in arriving at this conclusion.

These are the steps involved in building the chain to permit an inference about the truth or falsity of an hypothesis when we are doing experimental educational research. Obviously, not all educational research is experimental. Certainly there is very good educational research which deals with philosophical questions, but this falls outside the scope of this discussion.

The research process comes into better focus when we note a third aspect of this perspective on educational research, namely that there is a direct parallel between an experiment which is statistically evaluated and the problem solving that we do everyday. In the research situation, however, we are much more self-conscious about the way in which we perform each of the steps and about making sure that we have accounted for possible alternative explanations of the phenomenon that we are observing.

In both instances, we start with some sort of a hunch about what is true which leads us to make observations to ascertain the truth

or falsity of that hunch. The use of operational definitions and measures in educational research, however, is perhaps a more careful way of focusing our perceptions than we typically use. Similarly, the development of a sampling plan or an experimental design represents more careful attention to the matter of what and under what circumstances we observe than we typically apply to everyday problem solving. In everyday problem solving we typically apply some logic to the phenomena which we observe to determine the truth or falsity of our hunch. In the experimental situation, a combination of the statistics that we bring to bear on this situation, together with the experimental design, represents the application of logic. The logic being used, however, is a kind of mathematical logic applied to numerical data rather than the logic applied to a verbal description of situations. This application of logic permits us to evaluate the extent to which the observations support our hunch or hypothesis. But in experimental research the evaluation is given in numerical terms; in everyday problem solving, it is phrased in verbal terms.

The chain which we are attempting to build, then, is not an unfamiliar one. The chain is unfamiliar only when it is applied to a situation not typical of everyday life. As a chain is only as strong as its weakest link, so any argument is only as strong as each of its steps.

It should be noted that this chain of argumentation is a deductive argument. As Lord Hume pointed out many years ago, it is basically impossible to prove an inductive proposition by deductive argument. This gives us the fourth element in the perspective on the research process. With a deductive chain we cannot *prove* an inductive proposition. Thus it is better to view each experiment as a carefully evaluated instance in which a given proposition validly predicts the experiment's results, or is invalidated. An inductive proposition is true until we find an instance in which the hypothesis does not predict the experiment's results. Each experiment is an attempt to find another instance in which the hypothesis does predict accurately. If it should prove unsuccessful, the hunch or inductive proposition must be revised. An inductive proposition builds through the accumulation of a series of situations in which it has been found to predict successfully. As we demonstrate its predictive accuracy in each new instance, we tend to become more certain of its truth.

This in turn suggests a fifth aspect in our perspective on research. Since any instance can only be another confirming argument for an

inductive proposition, we see that a proposition is most useful if it grows out of a series of previously confirmed propositions or to put it another way, that it has a theoretical base.

One last aspect of the perspective on research may be gained by looking at this chain of reasoning. Almost any experiment has some flaw in it that might possibly invalidate the argument. This is particularly true of social science research. In reality it is almost never possible to build a completely tight chain of argumentation. Each chain is a compromise between what we can do and what we would ideally wish to do. Our statistical model almost never completely fits. Our experimental design is never completely tight. We are never sure that our sampling has not given us a biased sample. In each instance, we build the best possible chain of argument to show the truth or falsity of the proposition, but each design represents a compromise between the ideal and the possible. It is up to each person to evaluate the design and to determine whether indeed he will accept the evidence which stems from that compromise. In essence, he must examine the compromise to see whether it is satisfactory to him.

Let us review this perspective on research which we have attempted to sketch. We have noted that we are building a chain of argumentation in each experiment. This chain is parallel to the reasoning we use in everyday problem-solving. Each experiment is an attempt to determine whether the prediction of our hypothesis is invalid. We cannot by a single deductive chain of argumentation (which each experiment basically is) prove the hunch or hypothesis which we have. We can merely give another instance in which the hypothesis escapes invalidation. But even then our chain of argumentation is not completely tight. It is always a compromise between the ideal and that which it is realistic to expect in the situations in which we operate. We must examine each experiment to determine whether we are willing to accept the compromise which had to be made and to accept this as new evidence that the hypothesis has indeed escaped invalidation.

This is quite a different picture from the popular conception of the way in which research progresses, but it is nonetheless a realistic picture. When one considers the millions of explanations of phenomena that are possible but false, we realize that the world abounds with more false than true hypotheses. Thus the value of the process in winnowing out false hypotheses is certainly not to be discounted.

If this account is something less than perhaps we might hope, it

is all the more important to understand this process in this day and age, when we turn increasingly to research in the social sciences for help in answering our pressing problems. The social sciences tried things out and discarded old methods and explanatory concepts for years on an informal basis. Social science research is a kind of reasoning that brings added precision to an evaluation of those methods and concepts. It formalizes the criteria on which we decide whether to accept or to discard the propositions advanced. It makes public the criteria with clarity and also provides some basis for judging how well they are met. It helps cut through the bramble to a clearer decision. The tighter the research, the better its chain of argumentation, and the more carefully it is built, then the more value it has in providing a basis for decision making.

We have sketched some of the perspective on experimentation, and we have noted that each reader of research must examine the chain of experimentation for himself. This suggests that it would be well for us to examine the various steps in the chain a little more closely.

Let us start with hypotheses. We indicated that an hypothesis is a notion, a hunch, or a guess that something is true about our universe. We have already noted one thing which it is important to look for in an hypothesis or a hunch. Since each experiment is an instance in which an hypothesis is confirmed or disconfirmed, clearly one would be more likely to grant it credence if it is based upon previously confirmed hypotheses. This suggests that ultimately in the field of library science it is desirable to build a series of interrelated laws and propositions about what makes libraries more effective and what makes for better training of librarians. The building of laws or principles, and the testing of hypotheses which lead to new laws and new principles, is the line along which the most desirable kinds of hypotheses are to be found.

In what ways does the formulation of an hypothesis lead to the next step in the chain? The hypothesis contains the terms which must be operationally defined. These terms specify the kinds of phenomena that will be observed. The hypothesis indicates the nature of the relationship between variables to be observed. Thus it suggests the kind of situation in which they should be studied if we are to find such a relationship. This, of course, leads to a definition of experimental design. Finally, the generality with which the hypothesis is to be held suggests the nature of the sample on which we shall wish

to make our observations. Clear formulation of the hypothesis is thus a step with important implications for the rest of the chain of argument.

The terms in the hypothesis describe the phenomena which must be expressed in operational definitions. What is meant by an operational definition? The term "intelligence" illustrates one common example of an operational definition. Intelligence is nothing we can touch or feel or smell or in other ways subject to our senses. We can observe various acts which we define as exhibiting intelligence. Typically, we use a test situation for this purpose and we define the operations which lead to success on that examination as an operational definition of what we mean by intelligence.

Wherever we can, we quantify our operational definition. Thus we can evaluate how well the individual does on a test situation in terms of words; he did "well" or "poorly." But typically such descriptions do not convey as precisely as numbers how well the individual did. If in contrast we say that the individual had an intelligence quotient of 150, we know him to be an extremely unusual individual, in fact, we can tell in numerical terms how often such a score would typically occur. Quantification also permits us to apply statistical models, and so in experimentation we use quantification as often as possible to permit us to define our terms precisely and to make discriminations as exactly as possible.

We might parenthetically note that new fields of science typically start out with verbal description, moving to descriptive categories and finally to some sort of numerical scales as the field develops into a science. This development can be found historically in the natural sciences, and various social sciences are now in the process of transition. Although I am not acquainted with enough library research to know where your field lies today in this transitional process, it helps perhaps to realize that this kind of progression does exist.

Returning then to our chain of argument, we noted that the hypothesis suggested the generality with which the proposition should apply. But typically we do not investigate the phenomenon with all of the instances or all the people to whom it should apply. These people are thought of as the population, and we study the hypothesis as it applies to only a portion of these instances—a sample of the population. The means of selecting the sample is important, and is one step which has been carefully studied in the chain of argumentation. The manner of selecting a sample has implications for the kind

of statistics which can be used. We cannot go into a discussion of the ways in which samples are chosen here, but suffice it to say that this is one step in the chain which needs to be carefully handled.

The next step is that of experimental design, a key link in the chain. If you could change your library training curriculum so that you turned out better students, how would you be sure that it was this particular change which resulted in the production of these students? Perhaps it was that you had given more effort to the training; perhaps you had obtained new staff; perhaps you had provided additional facilities or texts. There are many possible explanations. Repeated observation of an experimental effect made in situations where we can be aware of other variables which might have caused that effect permit us to eliminate these variables as possibilities. But there are other ways we can rule them out. We may measure the effect of the contaminating variables, create a situation in which these contaminating variables do not occur, or arrange for the contamination to be held constant across the groups to be evaluated. The experimental design permits us to eliminate these alternative explanations and to control contaminating effects; it allows us insofar as possible, to isolate the effects of the experimental variables and measure them cleanly. Clearly this is a critical step in the successful forging of a chain of experimentation. We shall return to this step later and examine it in more detail, but let us proceed to the last step in the chain.

Finally, there is the statistical model which permits us to estimate the likelihood that the experimental effect observed is indeed an unusual one. Unusual in what sense? Unusual in the sense that it exceeds by some specified margin the results which we might expect from the fact that every sample chosen from a population will vary from every other sample and thus yield different experimental results. On occasion the particular sample chosen may have been a rarely occurring one which led to an unusual result; this is a chance that we must take in using statistics. Statistics merely tell us how unusual a particular result is. If the result is so unusual as to occur only rarely because of sampling fluctuation, we tend to believe that the instance which we are observing is likely due to something other than sampling fluctuation. If our chain of experimentation is tight, then we conclude that this is an instance in which the experimental effect that we are looking for has been observed.

In a sense we may think of the statistical model as a control for

the variability of sampling. It makes it possible for us to measure and estimate the effect of the sampling error. It permits us to estimate the likelihood that the situation which we observed was one which could be typically accounted for by sampling variability. If we find that we cannot account for the studied effect because of sampling error, if our experimental design is tight enough so that we cannot blame contaminating variables, if our operational definitions are acceptable, and if we have used a proper sample of the population to which we wish the generalization to apply, then we can assume that we have another instance in which the hypothesis that we are testing has been verified. This is the nature of the chain of reasoning that we build in the social sciences.

To this point we have tried to gain some perspective on the research process and to examine the steps in the chain of argumentation which permits the process to proceed. We have tried to indicate the place of experimental design in this chain of argumentation. It is a means of devising a situation in which the observations can be made so that various alternative explanations which might otherwise account for the phenomenon in question are ruled out. Now to look in greater detail at this particular step in the chain, to see how a design works. But to do this we need first of all to list some of the kinds of contaminating variables, some of the alternative explanations which might otherwise cause the phenomenon in question to occur but which are not the cause for which we are looking. The list that I shall use has been drawn from a chapter written by Donald Campbell and Julian Stanley entitled "Experimental Design" and published in the *Handbook of Research on Teaching* which was published in 1963.[1] I would urge that you consult this excellent source for a more complete description of the topics than I can possibly give here.

Let us begin our listing of rival hypotheses with an example. Suppose you are interested in the kind of reference use students make of a library as a result of the kind of teaching to which they are exposed. You are making observations in a number of schools in different communities, and in one of these communities the local television station happens to show a film on the use of the library. This, of course, is an event which is outside of your control but clearly would affect your observations. One might find that the children in this community were more expert in their reference work in the library than those in another community, and one might be led to

infer that this is due to the teaching in the school. This kind of event, which occurs during the experiment and which may affect the observations, is labeled "history" by Campbell and Stanley. It is an example of the kind of rival hypothesis that would account for the observations and lead one to believe that the experimental variable (in this case the kind of teaching) had an effect which it clearly did not have.

Let us examine some other contaminating variables or types of rival hypotheses. The effects of "maturation" processes within the persons observed which occur as a function of the passage of time also can produce effects which could be confused with experimental effect. Suppose one is studying story telling to very young children. Differences in age would result in differences in attention span which might frequently account for the way they respond to stories told them. Here the effect of maturation may be greater than the way the teacher tells the story or the kind of story that is told.

Another effect which is particularly important where we are dealing with two observations, observations before and after an experimental variable has been introduced, is named "testing" by Campbell and Stanley. Here we are concerned with the effects of taking a test on repeated testing; we are particularly concerned with what might be called the practice effect gained by the test taker. Clearly the second time a person takes a test he is more familar with it and is more likely to do better than the first time. Sometimes, also, he will have had a chance to discuss the test results and determine what is the "approved" or the "correct" answer and so the second testing reflects the "test wiseness" of the test taker rather than the effect of the experimental variable. Clearly this is another class of rival hypotheses that must be controlled.

"Instrumentation" is the class of rival hypotheses that arises because of changes in the way observers view a situation. Observers watching how businesslike students act in the library might use different standards at different times in observing the students. Interviewers trying to find out about reading habits might get different responses because of their own increased familiarity with the interview schedule over a period of time. Shifts in grading standards, learning how to administer a test, learning how to use an observation check list, these all constitute rival hypotheses to the main hypothesis, and they can result in the gathering of data which shows an effect like the experimental effect anticipated.

"Regression effect" is important where one deals with extreme groups. Suppose a group is picked because they did poorly on some test, a reading test, for instance. On retesting this group at a later date we can predict that they will have a higher average score than previously, not because of the effect of any treatment which might have intervened between testings, or because of the practice effect of the second testing. The change results from the imperfect correspondence of the score on one testing with scores in the second session. Unless we have a perfectly accurate measuring instrument, scores which are very high or very low may be expected to change on retest in the direction of the average scores of the population from which this group was taken. Our social science measures are almost never perfect. Thus, a poor group singled out on the basis of a test does better on retest after treatment, whereas a bright group may appear to have lost ground between pre- and post-test. Both findings are the result of regression effect, rather than of treatment effect. This is a very common finding in the literature, although it is rarely recognized as due to regression. The regression effect should, therefore, be anticipated wherever the selected groups are taken from the extremes of a distribution of scores and then retested on that same or a related measure to determine the effect of some experimental variables.

"Selection" is another source of rival hypotheses. Suppose you have a wonderful new information retrieval system for school libraries which you want to try out. Making it available on a voluntary basis you compare the themes of students who use the retrieval system with those who do not. Is there anything about those who volunteered which makes them special and which might have resulted in their doing better themes anyway? Clearly the different recruitment systems used in making up a sample may result in selecting people who are atypical in some way so that the effect observed is due to "selection" rather than to the experimental variable in question.

"Mortality," or the selective dropping of individuals from a group over the course of an experiment, is another source of rival hypotheses. Campbell and Stanley cite the fact that studies show freshmen women to be more beautiful than seniors. Does education decrease the pulchritude of college coeds? We would be unlikely to admit that this is the case. The rival hypothesis that a selected dropout exists because of marriage seems much more plausible.

Sometimes we have an interaction between one of the sources of rival hypotheses and the treatment effect that we are expecting to

produce. Such is the case when we pretest a group. We call this "interaction of testing and treatment." If we give a group a pretest, we focus their attention on the characteristics which we hope to change. We are thus more likely to cause increased change with respect to these variables. If you were to pretest students with respect to their knowledge of the Dewey Decimal system, when you later discuss the Dewey Decimal system in class, they are more likely to be alert for information about it than students who were not pretested. There is also likely to be greater retention of this information had you not raised the questions about the Dewey Decimal system as all. This is an example of the interaction between "testing" (in this case pretesting) and the treatment. There are other interaction effects, but we do not have time to discuss them here.

Let us discuss only one other source of rival hypotheses. The very fact that you are running an experiment is frequently the cause of change in the subjects. Many of you know of the Western Electric Company experiments which were done at the Hawthorne plant in Chicago in the 1920's. They were attempting to find how they could increase production. Whether they increased the lighting or decreased it, whether they improved the working arrangements or made them more awkward, whether they improved the ventilation or made it worse, they found that production went up because the workers felt they were special. The workers felt that they were part of an experimental group. Typically dubbed the "Hawthorne Effect," Campbell and Stanley use the name "reactive arrangements," to include the Hawthorne Effect and other aspects of an experimental setting to which the subjects might react. The artificiality of the experimental setting itself often results in an effect which is mistakenly taken to be the result of the experimental variable. The play acting, outguessing, up for inspection, "I am a guinea pig," or whatever other attitudes are the result of the experimental situation, are all included here.

This is by no means a complete list of all the sources of rival hypotheses. It is enough to give some idea of their nature, however, so that we can see how they are handled by an experimental design.

As the final step in this discussion, let us take a common experimental design and see the way in which this design provides some control on rival hypotheses. Let us study the typical control group design with which we are all familiar. In this design we have two groups to which individuals are randomly assigned. We observe these

two groups at the beginning of the experiment on those variables which we have operationally defined as resulting from the experimental treatment. One of the two groups chosen at random is subjected to some sort of experimental situation; the other is not. We then observe afterwards to determine what, if any, effect the treatment has had.

For example, a group of fifth grade children is assigned randomly to each of two classrooms. A classroom library is available to each of the classrooms, one arranged according to one classification system, the other arranged by a different system. We observe the children's skill in using these classroom libraries for reference work at the beginning of the year and at the end. Let us assume that the only training that these children have in using the library was given by an English teacher who serves both classes. Let us now look at the various sources of rival hypotheses that we have discussed.

Does this design control for "historical" events? In general, I think we can see that it does. Except for those events which might have occurred in one class but not in the other, "history" would be controlled. For instance, television programs instructing the children on use of the library would presumably be observed by as many fifth graders in one room as the other.

What about the effects of "maturation?" Presumably again these would be the same for both of the groups, since if we took a common pool of children and assigned them at random to these two groups, the effects of maturation or growth over the course of the experiment would be the same in the control as in the experimental group. The effects of "testing" similarly would be controlled in the sense that presumably background experience in testing would be equal for the two groups since we randomly assigned them to the control and experimental sections. Both groups would have the same pre- and post-observation experiences, so the effect of the second testing, as well as chances for them to discuss the test, would be comparable for both experimental and control groups.

The effects of "instrumentation" would also be typically controlled in that the observer's increased familiarity with the observation instrument would apply as well to the control group as to the experimental group. We should note, however, that one condition here is that the observer would not know which is the experimental and which is the control group. If the observer happens to be biased for or against the particular effect that one is seeking he might look

harder for it in the experimental group if he knew which group was which.

Even if these were extreme groups, which they happen not to be in our example, the effects of "regression" would be held constant across the groups, since the regression effect for two extreme groups randomly assigned to an experimental and a control session will be the same. The effects of "selection" would be equated between control and experimental groups, provided that the sample was chosen in one step and then assigned randomly to experimental and control groups. Similarly the effects of "mortality" would be the same in control and experimental groups, since presumably such effects as illness would cause children to drop equally out of both control and experimental groups. Should the experimental treatment prove distasteful, however, there might be an interaction between treatment and "mortality" which would cause an uncontrolled source of rival hypotheses.

What about another interaction, that between "testing" and treatment? This is an uncontrolled effect, for only in the experimental group do you have both the conditions of treatment and "testing." Thus this design does not control for the effect of inter-action between "testing" and treatment.

Does it control for "reactive effects?" This depends on how the control group is treated. If the control group thinks that they are special as much as the experimental group does, then we have this control between the two groups. If on the other hand, the experimental group realizes they are experimental but the control group does not, then "reactive effects" may also be a source of rival hypotheses.

Perhaps this is enough of an illustration to indicate the way in which experimental design can control for sources of change which might otherwise be confused with the one we wish to study. We could continue and discuss other experimental designs, and analyze them with respect to this incomplete catalog of rival hypotheses. The Campbell and Stanley chapter does this very well, examining a variety of such designs, as well as describing additional rival hypotheses.

In the short time available, hopefully you have gained some perspective on the research process, seen the function of experimental design in this process, learned some of the variables that can be controlled by experimental design, and seen how experimental design contributes to the chain of argumentation by controlling these variables. Hopefully also this will have stimulated enough interest in the process of experimental research that you will desire to study the

topic further, and bring experimental methods into your own field and further the stature of library science as a science.

Reference

1. Gage, Nathaniel, ed. *Handbook of Research on Teaching.* Chicago, Rand McNally, 1963.