Chapter 12

Research Design

Merle Canfield

Much of this discussion is taken from Campbell and Stanley (1966). Internal threats to validity attempt to answer the question "Are you sure that it was the identified independent variable(s) that resulted in the change in the dependent variable?" Or stated another way "Is there an alternative explanation for the resulting change in the dependent variable?" If there is any other possibility than the independent variable that is referred to as a "threat to internal validity." Campbell and Stanley offered 8 threats to internal validity ‑‑ other authors list anywhere between 8 and 10. The 8 of Campbell and Stanley are listed here.

 The internal threats to validity are as follows:

1. History of the subject. Any event that occurred other than the independent that may have produced the result in the independent variable. It may be anything that happened during the intervention or outside the intervention. This occurs when an event other than the treatment changes the subject on dimensions relevant to the dependent variable. For example, the dependent variable were depression and an event outside of treatment changed that relieved the depression the change attributed to the treatment would be invalid.

2. Maturation of the subject. This refers to any process that may occur over time. It differs from history in that it is systematic in terms of time ‑‑ growing older, getting smarter, getting bigger, using up resources, getting tired, habituating, getting more concerned, getting less concerned, and etc. If two groups of subjects, say first grade school children and third grade school children were assigned to brush their teeth with "Crest" and "Colgate" respectively for six months and assessed pre and post the "Crest" group might look bad because children loose their "baby teeth" about six years of age. Consequently, if number of teeth at pretest and number of teeth at posttest were used as the dependent measure erroneous conclusions would be drawn.

3. Instability of the subject. The state of the subject may change periodically (or randomly). For example, schizophrenic clients seem to have episodes where at times they are more lucid than others.

4. Testing effects. The effects of memory of the first test might have on taking the second test. The reduction of test anxiety ‑‑ or increasing test anxiety. The pretest can cue the subject to attend to the treatment. For example, if the pretest asks the subject about his or her level of stress and the treatment is designed to alleviate stress the subjects who were asked to assess their level of stress might be more attentive to the treatment and therefore be more influenced by the treatment.

5. Instrumentation. The measuring device may be unreliable. This is particularly true when raters are the measuring device. Raters might improve their accuracy over time because of practice or diminish accuracy because of boredom.

6. Regression artifacts. There is a tendency for extreme scores to move toward the mean over time. For example, when clients enter treatment they are probably "at their worst" and they are likely to "get better" regardless of what happens to them.

7. Selection. Subjects can be selected and a systematically biased manner. For example, if the subjects of two different psychiatric hospital wards are compared there is the possibility that the clients of one ward might be more chronic than the other ward.

8. Experimental mortality. When subjects withdraw from the treatment it may be that they do so because the treatment is not "working" for them. This will bias the study because the only the clients who remain will be tested at the posttest and the treatment was "working" for them.

9. Selection‑Maturation interaction.

External validity refers to the degree to which the results of the study can be generalized beyond the experimental situation. Factors that limit such generalizability are external threats to validity. Campbell and Stanley proposed 4 external threats to validity. Two more are added.

1. Reactive or interaction effect of testing and treatment. The pretest might que or sensitize the subject to the subsequent treatment.

2. Interaction of selection and treatment. The subjects may have been selected in such a way that the experimental subjects respond differently to the intervention than would the control subjects.

3. Reactivity to the intervention. The subjects may react to being in the experiment (either favorably or unfavorably).

4. Multiple treatment interference. When one part of the study effects another particularly when there are repeated interventions.

5. Irrelevant measure of outcome.

6. Irrelevant measure of treatment

Campbell, D. T., and Stanley, J. C. (1966) Experimental and Quasi‑Experimental Designs for Research. Chicago: Rand McNally and Co.

The following are various research designs that are intended to accomplish internal and external validity given the characteristics (limitations) of the experimental situation. The limitations of the experimental situation may be that it is unethical to randomize subjects to treatments, or that one cannot randomly assign subjects to diagnostic groups.

The format of the presentation is as follows:

Group 1: R O1 X1 O2

Group 2: R O1 X2 O2

Where:

Group 1 indicates group number 1

Group 2 indicates group number 2

R indicates that the subjects were randomly assigned to the group

O1 refers to observation 1 (or test 1 - in this instance pretest)

O2 refers to observation 2 (or test 2 - in this instance posttest)

X1 indicates that treatment number 1 was administered

X2 indicates that treatment number 2 was administered

moving from left to right indicates a passage of time -- for example randomization (R) occurs before the pretest (O1) and they both occur before treatment number 1 (X1) and so forth.

Os are dependent measures and Xs are independent measures

One-Group Pretest-Posttest Design

Group 1: O1 X1 O2

In this design there is only one group and the dependent measure is assessed before treatment (pretest) and after treatment (posttest). This design does not guard against any of the internal or external threats to validity. There design do little to protect against internal threats to validity. However, Campbell and Stanley suggest that selection and mortality could be protected.

Two-Group Pretest-Posttest Design Nonrandomized

Group 1: O1 X1 O2

Group 2: O1 X2 O2

This design could protect against history, maturation, testing, instrumentation, and regression. However, these are still at risk if the groups have different experiences in addition to the treatment. For example, if they are on different wards of a psychiatric hospital, or different classes in school.

Two-Group Pretest-Posttest Design Randomized

Group 1: R O1 X1 O2

Group 2: R O1 X2 O2

According to Campbell and Stanley this design protects against all threats to internal validity. It does not protect against the external threats to validity. Although Campbell and Stanley think there is a possibility that it could protect against the external threats of Interaction of Selection and Treatment and Reactive Arrangements. However, it would seem that mortality could be an issue if the treatment were noxious and caused subjects to withdraw.

Solomon Four-Group Design Randomized

Group 1: R O1 X1 O2

Group 2: R O1 O2

Group 3: R X1 O2

Group 3: R O2

Most authors agree that this design protects against all threats to internal validity. However, if one finds that more subjects withdraw from the treatment groups than the control groups suspicion is aroused.

One-Group Time Series Design

Group 1: O1 O2 O3 X1 O4 O5 O6

Campbell and Stanley suggest that this design protects against all internal threats to validity except Instrumentation. Other authors suggest that History and Mortality could be threatened.

Two-Group Time Series Design

Group 1: O1 O2 O3 X1 O4 O5 O6

Group 2: O1 O2 O3 X2 O4 O5 O6

Would seem to be better at protecting threats of History and Maturation than the One-Group Time Series Design. However, it would seem that both could be a threat.

Counterbalanced Design (Latin Square)

Group 1: X1 O1 X2 O2 X3 O3

Group 2: X2 O1 X3 O2 X1 O3

Group 3: X3 O2 X1 O2 X2 O3

Most authors agree that this design protects against all threats to internal validity except interactions between two or more threats.

Campbell and Stanley:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|   | History  | Maturation | Instability of the subject | Testing effects | Instrumentation | Regression artifacts | Selection | Experimental mortality | Selection‑Maturation interaction |
|   |   |   |   |   |   |   |   |   |   |
| One-Group Pretest-Posttest | - | - |   | - | - | ? | + | + | - |
| Two-Group Pretest-Posttest |   |   |   |   |   |   |   |   |   |
| (Nonrandomized) |   |   |   |   |   |   |   |   |   |
| Two-Group Pretest-Posttest  | + | + |   | + | + | + | + | + | + |
| (Randomized) |   |   |   |   |   |   |   |   |   |
| Solomon Four-Group | + | + |   | + | + | + | + | + | + |
| One-Group Time Series | - | + |   | + | ? | + | + | + | + |
| Counterbalanced |   | + | + |   | + | + | + | + | + |

 Philosophy of Science for the Psychosocial Sciences

Understanding human interaction is fraught with problems of epistemology ‑‑ "How do we know that, what we believe to be true, is true?" My interest in the philosophy of science is practical and is needed as a guide in selecting the proper research designs. This paper is about method; it is first philosophy of science and then a practical implementation of that philosophy.

 The purpose is to present a method of scientific investigation for the scientist studying human interaction. There are basically two questions to be answered when considering such research design. "Is there a relationship between what the hypothesis predicts and the resulting outcome data?" And "How can a researcher be sure that the variance of the outcome data is due to the variance of the treatment variable (theoretical construct)?" The first question deals with the fit of the theory to the actual data and the second deals with possible alternative explanations. The second question can be further elaborated with the following questions: (1) Will the relationship be contradicted by further evidence? (2) If repeated will the relationship hold? (2) Will the theory hold (relationships hold) in future situations?"

 Lachman (1960) states that the general objectives of science are to describe, comprehend, predict and control. More specifically to develop mathematical formulas to describe, comprehend, predict and control. The objectives are met by objectivity, caution, skepticism, parsimony, reduction of the complexity of mathematical formulas needed, and theory construction and utilization. Theory construction and utilization involves the following: completeness of formulation, coherence of the constituent components, simplicity, fecundity or fruitfulness, and precision of predictions. Consequently, the objective of science is to reduce the complexity of formulas needed to predict the outcome of a set of events in new situations.

 The goal of science is the search for universal patterns or laws. As Popper (1972) states, "The scientist will never let anything stop his search for laws..." (p. 247). Klenke, Hollinger and Kline (1980) state the same idea in a more subtle way "The terms 'definiteness' and 'precision' may be used in at least two related senses. First, they refer to the delimitation of our concepts and to the removal of ambiguity or vagueness. Second, they refer to a more rigid or exact formulation of laws. For example, 'It is more probable than not that X causes disease Y' is less desirable than 'The probability that X causes Y is 9.1.'" (p. 15). Neither Popper nor Klenke et. al. believe that psychology has any laws. Therefore, Popper would reject psychology as a science (he believes it will never become a science). While Klenke et. al. are willing to lower their standards and allow it to become a science.

 The position taken in this paper is that neither of these positions should be taken ‑‑ either rejecting the study of psychology as a science or lowering the standards so that psychology can meet the criterion of the new lowered standards. My position is that the highest standards of science should be maintained even though presently no laws exist. Lowering the standards retards the scientific development.

 The steps of a scientific method are presented and then these steps are compared to the requirements of science. Finally practical methods of putting these concepts into a practice of science are presented.

Method

 A six step method is presented and then the method is assessed in its ability to meet the requirements of science. The steps are as follows.

1. Develop a logically consistent theory, hypothesis, or theoretical construct that can be stated in such a way that a deduced formula or statement specifies the prediction of events.

2. Identify the scope of the events to be explained by the theory.

3. Describe a taxonomy of events (data) generated from the theoretical construct that is circumscribed by the scope. Include in this taxonomy any events that could falsify or test the theory if it were incorrect. Estimate the degree to which these variables represent the scope.

4. Describe a method to identify (measure) events (data) so that observations can be made reliably.

5. Test the predictability of these formulas (statements) against the standard of perfect predictability across situations within the scope.

6. Attempt a reduction of the number and complexity of the formulas (theoretical construct) for predicting events (testing fit). Compare the prediction of the formula to the prediction of other formulas. Seek alternative explanations.

 The first step of the method will probably not be contested by most science philosophers. Step #2 is a limiting step of the theory to a specific set of events. This step identifies those events that the theory should explain and likewise excludes those events that are outside the realm of explanation of the theory. Later it will be argued that theories with a broader scope should be selected over theories with a more limited scope. An example of limiting the scope is that the events to be explained are limited to say sociology or psychology. That is, even though the theory or laws should be universal (Step #5) the universality applies to the events within the scope specified.

 For example, a theory might be created to explain or predict cognitions rather than emotions. This is an arbitrary decision of the scientist. Even though there might be those who say that behavior cannot be separated from the biochemistry of the brain or that emotion cannot be separated from cognition the theorist has the option of setting the scope. At the same time it might be found later that another theory with the same number of statements or complexity of formulas that might explain the both sets of events. The theory that explains (predicts) the broadest set of events using the simplest formula should be selected as the reigning theory (Step #6).

 Step #3 might be more contested by some philosophers of science. The concept of Step #3 comes from Popper (1972) who states that the criterion of science is that theories should be stated in terms that are testable or falsifiable. He states, "One can sum up all this by saying that the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability." (p. 23).

 Step #4 refers to the reliability of observation. Basically this is a measurement and/or psychometric step.

 Step #5 follows from Popper's concept that science is the search for laws. It is most likely to draw criticism from the social scientists, however, I believe it to be the most crucial to a scientific method. Social scientists might not agree that the goal of social science is to search for laws ‑‑ that is to attempt to predict at the level of 100% correct predictions. I am not arguing that we can approach the level of 100% correct predictions only that the theory be tested against that standard. The goal is to search for universalities or laws even though we haven't found any yet.

 Step #6 closely follows Platt's (1964) modified version of Popper's corroboration concept (1972). This step could be thought of as a subset of Step #5. One theory could be compared to 100% and then a second theory compared to 100% and then these percentages compared. Step #6 is separated out because it becomes important in terms of method and the understanding of the relationship between philosophers of science. Step #6 is a more direct test of two competing theories. Platt (1964) argued that this is a major method in the development of science.

 Step #6 could be used to revise the theory and then start over with step #1. One now has a new theory to test. It should be recognized that this fits the philosophy of Kuhn (1970) but not the philosophy of Popper (1972). In Kuhn's (1970) rather sociological approach to the philosophy of science he describes 'normal science' as proceeding in small steps until there is a revolution. The changes made in Step #6 could be thought of as one of those small steps. This notion is in conflict with Popper (1972) in that he would have a theory be tested once and for all. If it is falsified the whole theory is falsified and it should be rejected and forgotten.

 Further it should be recognized the present method does not use the falsifiability of Popper (1972) but a modified version. In Popper's version a theory can only be falsified, never verified. Although he does have a concept of corroboration which is the absence of falsifiability. Platt (1964) modified the falsifiability concept to test competing theories. That is, the simplest theory, least falsified is selected as the reigning theory. It could be that two theories are equally good.

 The present method attempts to use the robustness of both of the methods of Popper and Platt. That is, that falsifiability is attained in the selection of variables (Step #3) and the competition of theories is in the result of the test (Steps #5 and #6). Further the present method tests the degree to which a theory is falsifiable. Some parts of the theory may be correct and therefore the theory might be revised by eliminating the false parts of the theory (Popper would reject this idea).

 Another aspect of this method is that step #1 is deductive while step #6 is inductive. Once the theory has been tested, step #6 helps to identify those parts of the theory might be improved. This can be done either by reducing parts of the formula (Ocums razor) or comparing it to another theory.

 Theories are selected based on: their scope (Step #2), the complexity of their statements or formulas (Step #1 and Step #6), and the level of their predictability (Step #5). The theory would be selected that had: simple formulas (or statements), broad scope and high predictability.

 What if a theory has simple formulas and high predictability but limited scope (for example studies with rats and the early learning theories might be classified into this category)? Should that theory be selected over one that has high predictability, complex formulas and broad scope. Weighing each of these dimension is not taken up at this point.

 These six steps are a description of the practice of science; included within them is the nature of science. All six steps are required for the practice of science. However, the question was raised by Popper (1972) of the criterion of science. What makes science different from other intellectual endeavors? Bartley (19‑‑) argued that the question of the criterion of science is relatively unimportant. It is left for the reader to decide its importance.

 Steps #2 and #5 combine to form the criterion of science. It is these two steps that separate science from art, law, or religion? That criteria is taken from Popper (1972) as the falsifiability of the theory by empirical evidence. The requirement of science is that a theory must be testable in the face of empirical data. At the same time the theory must be logically consistent (Step #1) but the criterion is falsifiability. Step #1 is not included in the criterion of science because it could also be performed by logisticians, mathematicians, theologians, and etc. This step does not discriminate scientists from other intellectuals. However, it is noted above that this step must be included as a part of the scientific process ‑‑ without it science is not being performed.

 There are other necessary features but it is this criterion of falsifiability that distinguishes it from other fields of thought and intellectual endeavor. There are two basic tenants of science: (1) don't give up the search for laws, and (2) the laws or theories proposed must generate formulas or concepts that would generate predictions that could be disproved by data. These two tenets are the criterion that separate science from other forms of intellectual endeavor but there are other necessary parts of the method that are shared with art, literature, poetry, religion, and law.

 The other steps could be part of other intellectual endeavors. They are necessary for science but they are not distinct from other methods. In fact they themselves could be researched and developed. For example, Kuhn's "philosophy of science" might more appropriately be labeled and as a sociological study of scientists. That is not devalue its usefulness but it is not a philosophy of science. It is not a statement of epistemology, method or a criterion for science. It is a hypothesis of the nature of scientists, and consequently, it is applied science itself.

Definition of Terms

 There has been a tendency in the behavioral sciences to describe Steps #2, #3 and #5 above with the term generalizability. That is, the term "generalizability" or the "generalization of results" has been used to replace or describe prediction of events across situations. If one word were to be used, predictability might be more descriptive and accurate.

 In order to met the requirement of generalizability the complete set of treatment and outcome variables need to be described. That means developing a taxonomy of the relevant set of variables.

 One of the problems in using the term prediction is that in some instances it has a limited scope. Authors like Pepper (1970) have argued against the use of predictability as the objective of science. However, it is not being used here in the sense he argued against. In this program a set of formulas are being used as the predictor. That is to say a theoretical construct is being used as the predictor so that it is the theory that is being tested. It is not merely a set of empirical variables. Further, the theoretical variable(s) is being tested within a set of other measured variables. As used here a theory should have postulates that predict events and these can be tested.

 It should be recognized that predictability is the more powerful form of generalizability. Predictability answers the question, "To what degree of probability can I predict what will happen in another situation?", while generalizability answers the question, "Will, what has an effect in this situation, have any (no determination of how much) effect in another situation?"

 The term explanation has a connotation of causality. At the same time it has implications beyond the empirical events at hand. It is often used to show the fit of the empirical events to the theory.

 At times throughout this paper the terms explanation, generalizability, and prediction are used interchangeably. In those cases where they are used interchangeably it is because of the references made to other investigators and should not lead to confusion. The term prediction is used as the preferred concept in this paper.

The Problem of Errors in Prediction

 The psychosocial sciences have a special problem not shared by other sciences. That is the problem of confounding variables which are difficult to isolate for laboratory study. For example, if the researcher is studying psychotherapy and identification of say five characteristics which make up the theory. bla bla

 The researcher might take these into the laboratory in some way

 It is important, however, to generalize back to the situation (psychotherapy). That is, using the theory to explain events to the general situation. One of two types of error can occur when using the theory to predict to general situations (Step #5): (1) the results indicate incorrect prediction when in fact the prediction was correct and (2) the results indicate correct prediction when in fact the prediction was incorrect. These are sometimes referred to Type I and Type II errors respectively. These can be categorized in the following way.

 A. Low prediction of outcome when the formula was correct

 1. Inaccurate predicting formula (theory)

 2. Errors in selecting variables (omission).

 3. Errors in measurement.

 a. predictor variables.

 b. control variables (if used)

 c. criterion variables

 B. High prediction of outcome when the formula was

 incorrect

 1. Criterion variable(s) do not contain falsifiable

 information

 2. Errors in selecting variables (wrong ones)

 2. Alternative predictor variables are not included.

 3. Predictor variables contain part of criterion

 variable.

 4. The criterion and predictor variables are mislabeled

 the criterion variable controls the predictor

 variable.

 Any scientific method must solve these problems. Two different methods are presented to solve the problem. These are then compared for their solution to the problem as well as their adherence to the other requirements of science as presented above.

Two Implementations of the Scientific Method

 Two models of implementing the scientific method are presented. They are then compared on their ability to meet the requirements of science and the method of dealing with the problem of errors in prediction.

 The major reason for this somewhat extensive excursion into the philosophy of science is that some form of randomized groups or trials is usually recommended as the best or only model of obtaining scientific integrity in social science. I would like to present and alternative model that I believe is a better fit to rigorous canons of science. The randomized groups method will be presented first and the multiple regression method follows.

 1. Randomized Groups Method of Psychosocial Science

 Psychosocial scientists have translated the philosophy of science into to a working model through the use of randomized groups or randomized trials. Student (190‑) first presented the method of comparing two groups. Fisher (19‑‑) expanded the model to more than two groups as well as testing for interactions. Solomon (1949) described the research design, and Campbell and associates (Campbell & Stanley, 1963; Cook & Campbell, 1979) identified where the errors that would occur when the model was put to use in the psychosocial sciences and prescribed the model of Solomon (1949) to avoid these errors. Errors would occur in that the theory would be accepted when in fact it was incorrect.

 Solomon (1949) proposed a procedure that used four different groups that would be needed in order to meet the requirements of a of a randomized experimental study. Campbell and associates (1963, 1979) identified the internal threats to validity that would be eliminated by this four group design. The design is presented schematically as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Time 1  | Time 2 | Time 3 |
| Group 1 | Test 1 | Treatment 1 | Test 2 |
| Group 2 | Test 3 | No Treatment | Test 4 |
| Group 3 |  | Treatment 2 | Test 5 |
| Group 4 |  | No Treatment | Test 6 |

 The implication underlying Campbell's internal and external forms of validity is that if these could be satisfied then the requirements of science or experimentation are satisfied. It is argued here that the scientific requirements are only partially met.

 The internal threats to validity are as follows:

 1. History of the subject.

 2. Maturation of the subject.

 3. Instability of the subject.

 4. Testing effects.

 5. Instrumentation

 6. Regression artifacts.

 7. Selection.

 8. Experimental mortality.

 9. Selection‑Maturation interaction.

 These internal threats can be summarized as follows. When there is only one group of subjects used in the experiment it is not known whether the treatment caused the effect or one of the above mentioned confounding variables produced the cause. For example, all of the children in a fourth grade class may have all had the same teacher in the third grade class that taught them arithmetic (long division) rather than the present one in the fourth grade class (history effect). In another example it might be that in a tooth brushing experiment of six year old children many of them lost their teeth. This could have been due to maturation rather than teeth brushing. The other internal threats are effects of the same confounding nature.

 These internal threats to validity are related to measurement, the correct selection of weights in the theoretical formula (or statement) and separation of variables from surrounding confounding variables. These are issues that are particularly problematic in the psychosocial science. The are slippery and mean.

 The external threats to validity are:

 1. Interaction effects of testing.

 2. Interaction of selection and experimental treatment

 (unrepresentativeness).

 3. Reactive effects of experimental arrangements.

 4. Multiple‑treatment interference.

 5. Irrelevant measures.

 6. Irrelevant assessment of treatments.

 These threats to validity can be rephrased in the following manner. In order to generalize to future situations one needs to know the degree to which existing extraneous variables will affect the independent variables (predicting formula as measured by some set of variables) and dependent variables (outcome of the experiment) in those future situations. The following assumptions are made by the randomized groups method. The extraneous variables are randomly correlated with the experimental variable and the control variable. The extraneous variables are not correlated or that they are randomly correlated with other extraneous variables. It is further assumed that randomization will result in the sample representing the situation to which one wishes to generalize.

 Deficits of the Design

 It cannot be assumed that the extraneous variables are: (1) randomly correlated with the control and experimental conditions: (2) not systematically intercorrelated; and (3) that they are representative of the situation to which one wishes to generalize. In fact is most likely that such correlations would occur.

 It is somewhat surprising that a number of advocates of the design also state that the design does not solve the external validity problem. In reference to the two‑group design Crano and Brewer (1973) state: "No matter how good the experimental design, the question of external validity is still unanswerable. It is good to attempt to approach the ideal of generalizability, but foolish to expect its attainment through the use of experimental techniques that simply were not designed to provide such assurances." As Campbell pointed out the model does a good job of eliminating the errors of internal validity but does nothing about the errors of external validity. Campbell (1975) referred to the errors of generalizability as "threats to external validity". He states (Campbell, 1975): "These threats apply equally to true experiments and quasi‑experiments."

 It is disquieting when the advocates of the preferred method indicate the method does not meet the requirements of science. That is, that generalizability is the stated goal of science according to Campbell et. al. and yet even the most stringent method proposed by them (the Solomon four group design) does not attain such generalizability. Further, it is postulated here that the method of attaining internal validity (randomization) results in this lack of generalizability. A strange state of affairs.

 The randomized groups design fails to meet the test of generalizability and, consequently, the more stringent requirement of predictability. These external threats are related to universality of the hypothesis. These problems arise basically because of the method used to solve the internal validity problem. If the internal validity problem could be solved by a method other than randomization then the external problem might not exist. Confounding of variables still remains unsolved and problematic and, consequently, solving the internal validity problem by randomizing solves nothing. The reason that randomization causes problems for external validity is that randomization does not separate or eliminate the effects of the confounding variables from any variables except the treatment variables. If variables could be separated properly and measured then the external relationships could be solved directly.

 2. Multiple Regression Method of implementing Psychosocial Science

 The second method that social scientists have used to implement the scientific method is to solve the "prediction error" problem with methods other than randomization. This will be referred to as the multiple regression model (RM). The method itself goes much beyond the statistical model implied. It includes taxonomizing, describing, and testing the independent and dependent variables. A major effort is made to include variables which could falsify the theory.

 Rather than randomize in order to obtain interval validity the researcher using this method identifies alternative hypothesis and moves toward what Platt refers to as strong inference. The basic idea of this method is to test one variable against another; to search for the falsifiable variable; to include it in the study; to search for the underlying unknown variable.

 For example, the researcher might include variables of history, maturity, and etc. In the early stages of research in an area (specified by the scope) variables will multiply but as they become known (because of describing them not randomizing them) the researcher will be in position to choose those which may effect the outcome (alternative explanations).

 While the statistical prototype of the randomized groups (RG) method is the t‑test the statistical prototype for this design is the correlation. The statistical model is the multiple regression (MR).

 The stringent requirement of science is that a scientific theory be tested against the level of perfect predictability which the MR model affords. This is a way of estimating both its internal and external validity. However, one can never be sure of either but its final estimate is known. A theory, formula, or hypotheses is presented to be tested against the perfect predictability. Not only is it more "hard nosed" science but it also allows the researcher to choose broader more complex problems of human intercourse; to study the complexity of many interesting variables in vivo; in vitro; in toto. [enter Aronson]

 Structural equation modelling and confirmatory factory analysis are methods for testing hypotheses of covariance structures (Joreskog, 1974; 1979). The methods compare goodness of fit of hypothesized models of relationships among variables to observed covariances (or correlations) among a set of variables. The analysis of covariance structures is a very general approach for correlational data, and includes many typical multivariate procedures (e.g., multiple regression, path analysis, factor analysis, analysis of variance and covariance, canonical correlation) as special cases of the general model. Unlike many of the correlational multivariate methods, however, the analysis of covariance structures allows an a priori specification of variable relationships. Thus, some relationships may be set to zero (or some other value) or constrained to be numerically equal across variables or groups. Systematic hypothesis‑testing is possible by comparing the goodness of fit between alternative models that reflect different hypotheses about the relationships (Step #5).

 An example might be helpful to show how the method deals with the error of prediction problem. Comprehension and prediction are increased when the number and complexity of mathematical formulas are decreased. Suppose that over a 75 year period a correlation of 0.56 was found between the number of babies born each year and the amount of money spent on highway construction. Further assume a negative correlation of ‑.36 was found between the number of goats sold and babies born, and the number of goats sold correlated with amount of money spent on highway construction was ‑.48. Such a set of correlations does not seem to "make sense" until it is found that they are all related to the "state of the economy". Our understanding and predictive capabilities are increased by "reducing" the number of variables. The two‑group design does not foster the search of these extraneous but related variables. Nor does it contain within it a method for reducing the number of "explaining" variables.

 Deficit of the Design

 The lack of internal validity remains as a major flaw in this design. The researcher cannot be sure whether the resulting effect was due to the independent variable (statement or formula used for prediction) or whether it might have resulted from some confounding variable. The example given above was the history and maturation.

Comparison of the two Methods

 How do the two models meet the requirements of a scientific methodology? Table I presents a grid with the two methods at the tops of the columns and the characteristics of the scientific method down the side with statements

 Table I. Comparison of the method of Randomized Groups and Multivariate Designs.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Randomized Multiple

Groups Regression

Designs Designs

Internal Validity yes weak

External Validity no weak

Predictability no yes

(Generalizability)

Confounding Variables no yes

test

Alternative Hypothesis weak yes

test

 The trade off of the two methods is the following: the MR model has weak internal validity for stronger external validity while the RG model has strong internal validity with unknown external validity. The MR model allows two important tests: (1) comparison to alternative hypothesis and (2) comparison to 100% predictability. Both of these are important to the relationship to the criterion of science.

 The randomized groups method is strong in the category of internal validity while sacrificing all of the other requirements of science except the alternative hypothesis test where it is weak. While the multiple regression (MR) design is weak in both internal validity and external validity is strong in all other areas. Further, the MR model fosters developments so that in the long run it will overcome both internal and external validity.

 It is the task of science to identify and separate out the confounding variables not randomize them so that we keep ourselves ignorant. Not only does it require information about which variables they are but also their effect on the outcome. How much of the total variance of the outcome do they impact? It is knowledge of these effects that will first help us to eliminate internal validity errors and later approach perfect predictability. This method fosters the development of science in that it impacts on all six of the step of the scientific method.

 The following is an assessment of the two methods in their ability to meet the six steps of scientific investigation.

 Step #1. When multiple regression is used a model of the formula can be generated by using the beta weights. In the same way weights can be used in logit models. These weights can be assessed directly in the empirical test. However, with the randomized groups method the amount of the independent variable is rarely measured and consequently the amount is unknown.

 Step #2. There is nothing in either of the methods that is particularly helpful or lacking in this step.

 Step #3. Since all relevant variables must be included in the final test of the MR design they are more likely to be known to the investigator. Furthermore, their relationships are tested so that past experiments will have indicated their hierarchical relationship so that only relevant variables need to be selected.

 Step #4. Since errors of measurement will reduce predictability measurement will be improved in such a program. This will further improve precision. In the RG method the predictor variables (independent variables) are rarely measured or assessed.

 Step #5. This is the real test of the theory. The two‑group design does not meet the requirements of science because it: (1) does not meet the minimal requirement of predictability across situations; (2) does not meet the more stringent requirement of comparing predictability against 100% of reducing the number and complexity of formulas. Although the latter could be done using the two‑group design ‑ the design does not foster the reduction.

 The MR model identifies and measures amount of treatment, control and outcome variables, followed by an assessment of the level of prediction. This is accomplished in the MR model and not tested in the RG model.

 Step #6. This step is concerned with alternative and/or simpler explanations. Both models accomplish this step. However, in the RG model and amount of difference between two competing theories is unknown, the difference is known in the MR model. Since the difference is known in the MR model, and since the variables are all assessed there is a check on internal and external validity in this step.

 Furthermore, the model simplification is built into the statistics of the MR model. For example, in both multiple regression and LISREL there are methods of adding and eliminating variables to note their effect. Consequently, simpler formulas can easily be assessed. If fact they are assessed as part of the experimental process.