

## Basic Principles in the Design and Evaluation of Experiments

Robert M. McFatter  
 University of Louisiana  
 Lafayette

The interpretation of statistical analyses of data derived from a study may be carried out at several levels. At one level, the results may be interpreted simply as evidence addressing questions such as whether there are differences among the means of several populations or whether the variables appear to demonstrate various kinds of associations among themselves. For this level of interpretation, *statistical* assumptions (e.g., normality, homogeneity of variance, etc.) may well be the primary considerations. Often, however, the analyst wishes to draw conclusions about causal processes, mechanisms of action, etc. Interest in these kinds of inferences is common in scientific investigations. It is important in thinking about the interpretation of any set of data to consider the nature of the study's design and the possible limitations that design may place on inferences drawn from the data.

### Experiments vs. Quasi-experimental Studies

It is common to distinguish between *randomized experiments* (sometimes called 'true experiments') and *quasi-experimental* studies (sometimes called 'correlational studies'). It is important to realize that, although the *statistical* analyses of these two kinds of studies can often be *identical* to each other, the substantive interpretation of the two kinds of study *must* be carefully distinguished. The two key features of a randomized experiment are (1) random assignment of subjects to conditions (NB. this has nothing to do with random sampling!), and (2) controlled manipulation of the independent variable. In a quasi-experiment, by contrast, the experimenter does not randomly assign subjects to different groups nor even manipulate which group a subject is in, but rather simply observes which group a subject is in. The distinction may be illustrated with a simple example. Suppose an investigator is interested in the effect of vitamin E on diastolic blood pressure. Two studies with 24 subjects each are carried out which just happen, amazingly enough, to produce identical data as given below:

	Dosage of Vitamin E		
	0 IUs/day	200 IUs/day	400 IUs/day
	95	91	73
	79	72	81
	100	86	66
	85	80	81
	96	85	63
	91	75	60
	82	90	75
	100	69	77
Mean	91	81	72

The two studies differ, however, in their design. In the first study, a random sample of eight individuals was found who reported taking no supplemental vitamin E each day for the last 6 months. Their scores are in the first column. Two additional random samples of eight individuals were found who reported taking 200 and 400 IUs/day, respectively, of supplemental vitamin E for the last 6 months. Their diastolic scores are reported in the last two columns.

In the second study, the investigator took a random sample of 24 subjects from the same general population as in the first study, but randomly assigned eight of the subjects to receive either a placebo, 200, or 400 IUs/day of supplemental vitamin E for 6 months. After 6 months diastolic pressures were obtained for each subject as shown in the table.

Statistical analyses were carried out on the data from each study (one-way ANOVA is the appropriate analysis), and, of course, the results were identical. The dosage effect was highly significant,  $F(2, 21) = 10.79, p = .0006$ , indicating that the population mean diastolic blood pressures for the three groups were different from one another in both studies. Thus, the *statistical* conclusions from the two studies are identical. However, the *interpretations* of the results in the two studies are quite different. The first study is a quasi-experimental or correlational study, whereas the second is a randomized experiment. The conclusion one is entitled to draw from the first study is that diastolic blood pressure *is associated with* supplemental vitamin E dosage in such a way that individuals who take more vitamin E on average have

lower pressures. One is *not* entitled to conclude on the basis of this information alone, however, that taking supplemental vitamin E has an effect on blood pressure. The reason is that the population of individuals who take 400 IUs/day very likely differs in many ways other than vitamin E intake from the population of individuals who take no supplemental vitamin E. For example, they might be more health conscious, more intelligent, or have more pets! These other differences might well be the cause of the observed blood pressure differences. Although the differences in diastolic pressure found *may be* the result of differences in vitamin E intake, there are many other *plausible alternative explanations* for those differences.

One is entitled to draw much stronger conclusions from the second study. It is crucial to understand the role that random assignment plays here. Randomly assigning subjects to conditions insures that if an infinite number of subjects were assigned to each condition the three populations would be identical on average on *all variables* except vitamin E intake. '*All variables*' here would include health consciousness, intelligence, number of pets, and any other variables that could potentially influence blood pressure. So if one concludes that the three populations' mean blood pressures differ (which is what the ANOVA tests) then the only explanation of that difference is vitamin E intake. Therefore, we conclude that vitamin E dosage differences *caused* the differences in diastolic pressure. Random assignment is an extremely important and powerful tool in experimental design.

It is a common error for investigators and readers to draw unwarranted causal inferences from quasi-experiments. An awareness of the subtle ways in which this error may manifest itself is crucial in designing and evaluating scientific studies. Nonetheless, it would be going too far to suggest that it is never reasonable to draw causal inferences apart from evidence obtained from a randomized experiment. Most people draw sensible causal inferences about a host of variables every day without doing a single randomized experiment! The key criterion for making justifiable causal inferences has to do with the degree to which *plausible alternative explanations* of the observed effect can be ruled out. There is no single universal method for accomplishing this, and often a variety of considerations will play a role in making a convincing case for a causal inference. Randomized experiments are simply a very powerful way of ruling out a large number of plausible

alternative explanations all at once. It is worth noticing that it is not the kind of numbers, variables, or type of statistical analysis that determines whether causal inferences are justified, but rather how well one is able to rule out plausible alternative explanations.

### Experiments as Attempts to Explain Variation in a Dependent Variable

In the design and evaluation of experiments it is useful to think of an experiment as an attempt to explain variation (variance) in a dependent variable. The terms *independent variable* and *dependent variable* are used to refer to variables that play important roles in the study. Generally speaking, one wishes to account for or explain variation in some variable. This variable is called the *dependent variable*. Often one is interested in the influence or effect of some other variable on the dependent variable. This other variable would be called an *independent variable*.

Notice the causal language here. Variations in the independent variable *cause* or *produce* variations in the dependent variable. In the experiment described above, vitamin E dosage would be the (single) independent variable (with 3 levels: Placebo, 200 IUs/day, and 400 IUs/day), and diastolic blood pressure would be the dependent variable. Because the terms 'independent' and 'dependent' variable have such causal overtones many authors restrict the usage of the terms to randomized experiments or, at the very least, experiments involving controlled manipulation of the independent variable. It is only from these kinds of experiments that strong causal inferences might be derived. These authors, therefore, would *not* use the term 'independent variable' to refer to vitamin E dosage in the *quasi-experiment* described above, but would prefer some other term like 'predictor variable' to avoid the unwarranted causal implications associated with the term 'independent variable.'

Other authors are not so fastidious in their use of the terms and happily use 'independent variable' to refer to any kind of predictor variable even in quasi-experiments. No doubt this usage derives from a recognition that the statistical models for the two kinds of study are identical. I generally try to use 'predictor' and 'criterion' to refer to the variables in a quasi-experiment that would be analogous to 'independent' and 'dependent' variables in a randomized experiment. The important point, however, is that regardless of which terms are used,

appropriate attention must be paid to the kinds of inferences that are justified, given the design of the study.

In both kinds of study it is sensible to consider the statistical analysis as an attempt to explain or account for variation in the dependent or criterion variable by measuring the degree to which that variation is related to or associated with variation in the independent or predictor variable. In the example above we may say that individuals vary in their diastolic blood pressures. The question we ask in the statistical analysis is, "How much of that variation is caused by variations in vitamin E intake?" (in the case of the randomized experiment); or "How much of that variation is predictable from vitamin E intake?" (in the case of the quasi-experiment). In what follows I will focus on the case of the randomized experiment, but the ideas may be applied as well to the quasi-experimental case.

Because variation in the dependent variable is measured by the variance of that variable, it is common to address the question of how much of the variation in the dependent variable is caused by the independent variable by breaking down or partitioning the variance of the dependent variable into pieces that reflect different causes. One useful partition is as follows:

*Primary variance:* Variance in the dependent variable that is due to, or caused by, variations in the levels of the independent variable.

*Secondary variance:* Variance in the dependent variable that is *systematically associated with* some potentially identifiable variable or variables other than the independent variable.

*Error variance:* Variance in the dependent variable that is unsystematic, unpredictable or 'random.'

Many of the issues in experimental design involve controlling and manipulating these different sources of variance in the dependent variable. Generally speaking, a good experimental design will, among other things, tend to *maximize primary variance*, *minimize error variance* and *control secondary variance*.

The idea in maximizing primary variance is that in designing the experiment the experimenter chooses levels of the independent variable that are expected to produce as large an effect on the dependent variable as

possible. The reason for doing this, of course, is that the larger the effect size in an experiment the more likely it is that the hypothesis test in the statistical analysis will be able to detect it. In other words, the larger the effect size, the more powerful the hypothesis test will be (i.e., the lower will be the probability of a Type II error). By 'effect size' we mean, in general terms, how large the differences are among the dependent variable means for the various levels of the independent variable. In the vitamin E experiment, maximizing primary variance would involve choosing dosages of vitamin E for the study that would be expected to produce the largest differences in diastolic pressure.

Minimizing error variance and controlling secondary variance often go hand in hand because one way of reducing error variance is to control secondary variance in the dependent variable. Error variance can be thought of as the 'noise' variation in the dependent variable. We are interested in detecting a 'signal' (differences in scores on the dependent variable that are caused by variations in the independent variable), but that signal is disguised by unsystematic or unpredictable noise variations in the dependent variable. To the extent that noise variations in the dependent variable can be reduced, the signal will become more detectable.

In an experiment that is analyzed using ANOVA, error variance is measured by the denominator of the *F*-ratio. In an ordinary one-way between subjects ANOVA the  $MS_{\text{within}}$  would be the estimate of the error variance in the experiment. Recall that the  $MS_{\text{within}}$  here would be the average of all the variances of the dependent variable within the groups. Thus, it would be a measure of how variable the scores in each group are around their group mean. Note that this is unpredictable variation. If all one knows is which group a subject is in, the best prediction of that subject's score on the dependent variable is the mean for the particular group. That prediction would be identical for every subject in a particular group. Thus, variations *within* the group around the group mean score represent unpredictable or noise variations.

It is important to consider what the sources of this unpredictable variation might be. One source of unpredictable variation might be what are generally referred to as individual differences among subjects on some variable that is related to scores on the dependent variable. For example, in the vitamin E experiment one source of variation in scores within treatment groups might be simply the age of subjects.

It is well known that blood pressure tends to increase with age. If subjects of different ages are included in each group some of the error variance that the  $MS_{\text{within}}$  measures will be due to differences in the age of subjects. Notice that if age is not measured or controlled in some way in this experiment variance in diastolic pressure due to age will simply show up in the error term as unpredictable variance. However, variance in diastolic pressure due to age is not intrinsically unpredictable—it is, to a certain extent at least, *systematic* and can be treated as a form of secondary variance. The idea is to *control* secondary variance and in so doing remove that source of variance from the error term. Removing this source of variance from the denominator of the  $F$ -ratio makes the  $F$  larger, i.e., increases the likelihood that real treatment effects will be detected.

How can this kind of control of secondary variance be carried out? There are basically four methods of controlling secondary variance. They are

- 1) randomization (i.e., random assignment of subjects to conditions)
- 2) holding secondary variables constant
- 3) systematic variation of secondary variables
- 4) statistical adjustment for secondary variables.

#### *Randomization*

The first method, randomization, does not really reduce the error variance in the experiment, but, as we have already seen, it simply insures that the populations in each treatment group are identical on all secondary variables, thus removing any systematic biases in the group means that would be due to secondary variables. For example, in the vitamin E experiment randomization would insure (among other things) that all ages would be equally represented in each treatment group.

#### *Holding Secondary Variables Constant*

The second method, holding secondary variables constant, reduces error variance by simply eliminating any variation in the secondary variable that could produce variations in the dependent variable. In the vitamin E experiment one might include in the experiment only subjects from a very narrow age range, say, age 50-55. Without the variations in diastolic pressure due to age, the  $MS_{\text{within}}$  would, of course, be smaller (leading to a larger  $F$ ) than it would be if individuals of widely varying ages comprised each group.

#### *Systematic Variation of Secondary Variables*

The third method is a generalization of the second. A good example of the use of systematic variation of secondary variables would be a randomized blocks design. In this design several groups of different subjects are first formed on the basis of their scores on a concomitant or “blocking” variable. This blocking variable is the secondary variable or a variable that measures or is correlated with secondary variation in the dependent variable. Subjects within each level of the blocking factor are then randomly assigned to be in one of the levels of the primary independent variable. A score on the dependent variable is obtained for each subject. Generally, the blocking factor one chooses is one that is expected to be highly correlated with scores on the dependent variable.

In the vitamin E experiment age would be a reasonable blocking factor. An experimenter might first block subjects into, say, four age groups, and then within each group randomly assign subjects to receive one of the three dosages of vitamin E. A two factor between subjects ANOVA with age and dosage as the independent variables and diastolic pressure as the dependent variable would be a natural analysis.

It is worth noticing how such an analysis controlling the secondary variance due to age reduces the error variance in the design. If one did not control for age by blocking on it first, the appropriate analysis would be simply a one-way ANOVA with three treatment groups. The error term for that analysis, the  $MS_{\text{within}}$ , which would be simply the average of the variances within the three treatment groups, would include variation in diastolic pressure that was in part due to age differences among subjects within each treatment group. In the two-way ANOVA with age as a blocking factor the error term would still be the  $MS_{\text{within}}$ , but this time it would be the average of the variances within each of the 12 (i.e.,  $3 \times 4$ ) age  $\times$  treatment cells. Because each of these cells is relatively homogeneous with respect to age, variance in diastolic pressure due to age would be no longer part of the error term. This reduced  $MS_{\text{within}}$  would lead to a larger  $F$ -ratio and an increased chance of detecting any real treatment effect.

Another way of thinking about what is going on here is to recognize that the treatment group means in the one-way analysis (ignoring age) would be expected to be the same as the treatment main effect means in the two-way analysis. Therefore, the  $SS_{\text{Between}}$  in the one-way ANOVA would be identical to the

treatment main effect  $SS$  in the two-way ANOVA. With the same number of subjects in both analyses the  $SS_{\text{Total}}$  would also be identical in the two analyses. This means that the  $SS_{\text{Within}}$  of the one-way analysis would equal the sum of the  $SS_{\text{Age}}$ ,  $SS_{\text{Treatment} \times \text{Age}}$ , and the  $SS_{\text{Within}}$  of the two-way analysis. Thus, in the two-way ANOVA secondary variance due to age ( $SS_{\text{Age}}$ ) or the differential effects of the treatment on different age groups ( $SS_{\text{Treatment} \times \text{Age}}$ ) have been partitioned out (i.e., removed) from the error term.

### *Statistical Adjustment for Secondary Variables*

The fourth method of controlling secondary variance involves including measures of secondary variables as additional predictors in a regression analysis type statistical model of the experimental design. Models like this are sometimes referred to as analysis of covariance models. This approach to controlling secondary variance is powerful and important, but beyond the scope of the present discussion.

It should also be pointed out that error variance is also affected by the carefulness and precision of the experimental procedures used. In addition, the reliability and validity of the dependent measure are certainly factors that affect the size of the error variance. Minimizing error variance requires careful attention to any factor that produces 'noise' in the dependent variable.

## **Theory and Hypothesis Testing in Experiments**

Most, if not all, experiments and quasi-experiments are carried out with the idea in mind of testing a theory or hypotheses derived from a theory. Even exploratory studies are usually carried out with at least some tentative hypotheses in mind. The most elegant scientific studies are often those which derive rival predictions from more than one theory and generate empirical data that allow one to decide which theory better fits the evidence. At the end of the analysis the investigator will generally want to draw a conclusion about how the data support (or fail to support) a particular theory or hypothesis.

For example, in the vitamin E experiment, the investigator might wish at the end of the analysis to be able to claim that the pattern of results found in his data supports the hypothesis that increased dosages of vitamin E produce lower diastolic blood pressures. In order to make this kind of argument convincing,

however, the experiment must be designed in such a way that critical readers find it very difficult to come up with plausible alternative explanations of the results. That is, the results should be unambiguous in their interpretation. Designing experiments that are convincing in this way can be quite challenging. As Maxwell and Delaney (1990) point out:

"Indeed, part of the art of experimental design has to do with devising control conditions for which the theory of interest would make a different prediction than would a plausible rival hypothesis. (For example, the rival: 'The deficit is due simply to the operation not the brain area destroyed' is discounted by showing no deficit in a sham surgery condition.) If the rival hypothesis is false, part of the credo of science is that with sufficient investigation it will be ultimately discovered. As Kepler wrote regarding rivals to the Copernican hypothesis that made some correct predictions,

And just as in the proverb liars are cautioned to remember what they have said, so here false hypotheses which together produce the truth by chance, do not, in the course of a demonstration in which they have been applied to many different matters, retain this habit of yielding the truth, but betray themselves (Kepler, 1601).

Although in principle an infinite number of alternative hypotheses always remain, it is of little concern if no *plausible* ones can be specified." (p. 18)

A focus on designing experiments in such a way as to eliminate plausible alternative explanations leads naturally to thinking about the validity of the inferences that may be drawn from an experimental result. Actually, it is useful to consider the evaluation of experiments by using concepts that may be more familiar as general psychometric properties commonly used to evaluate psychological tests or measures.

## **Reliability and Validity of Experiments**

The two psychometric concepts commonly used to evaluate tests are *reliability* and *validity*. The reliability of a test generally refers to the stability or consistency of measurement provided by the instrument. There are a number of ways commonly used to assess the reliability of a test: test-retest, equivalent forms, split-halves, coefficient alpha, etc.

The validity of a test refers to the extent to which

the test measures what it is intended to measure. It is common to distinguish different ways of assessing the validity of a test: criterion-related validity (including predictive and concurrent validity as subtypes), content validity (with face validity and intrinsic validity as subtypes), and construct validity.

The psychometric concepts of reliability and validity may be usefully applied to the evaluation of experiments as well as tests.

#### *Reliability of an Experiment*

The reliability of an experimental result would refer to the stability or consistency of that result. That is, if the experiment were repeated under similar conditions, would a similar result be obtained? There are two general ways the reliability of an experimental result may be evaluated. The statistical significance test commonly employed in evaluating experimental results is, in fact, a way of getting at one aspect of the reliability of a result. The  $p$ -level associated with a significance test of a treatment effect reflects the probability that if the null hypothesis were true (i.e., if there were no real population treatment effect) one would obtain a sample treatment effect as large as, or larger than, the one found. A small  $p$ -level thus indicates that a sample treatment effect as large as the one obtained in the experiment would be quite unusual if there were really no true population treatment effect. To the extent that a low  $p$ -level reflects a real treatment effect, then, one would expect other similar experiments to be likely to show similar treatment effects.

A more direct, and in many ways more powerful, argument for the reliability of an experimental result comes from replications of the experiment (preferably by other investigators) in which similar patterns of results are found. Replication is a fundamentally important part of the scientific method, and there is really no substitute for it.

#### *Validity of an Experiment*

The concept of validity as applied to an experiment represents the extent to which the design, execution and interpretation of the experiment actually justify the conclusions the investigator wishes to draw from the experiment. It has become common since the influential work of Cook and Campbell (1979) to identify four types of validity to be considered in evaluating experiments. They are *internal validity*, *external validity*, *construct validity*, and *statistical validity*.

*Internal validity.* The internal validity of an experiment refers to the extent to which the design and execution of the experiment allow the conclusion that the treatment (rather than some other factor) caused the differences in the dependent variable. As we argued earlier, randomized experiments generally allow one to draw stronger causal inferences than do quasi-experiments because the random assignment of subjects to conditions allows one to rule out a host of plausible alternative explanations of the differences found. Consequently, randomized experiments tend to have stronger internal validity than quasi-experiments.

There are a number of conditions that constitute 'threats' to the internal validity of an experiment. These threats represent common problems that result in causal inferences from the experiment being unconvincing or misleading. These threats include the problems of confounded variables, differential experimental mortality (different kinds of subjects drop out of the experimental conditions), artifacts, carryover effects in repeated measures designs, communication between subjects in the course of the experiment, as well as others. We will return to discuss some of these threats to validity in more detail after describing the other types of validity.

*External validity.* The external validity of an experiment refers to the degree to which the results of the experiment may be *generalized* to populations, settings, treatment variables, and measurement variables *other than the ones in the experiment itself*.

For example, suppose one conducts an experiment looking at the effects of different fonts on the readability of text. In order to maximize the internal validity of the experiment conditions are carefully controlled, and subjects drawn from a random sample of college students are randomly assigned to different font conditions. A standard passage of text is presented in the different fonts on a computer screen in the lab. The text in all conditions is always presented as white text on a black background. The sizes of the fonts are the same in all conditions, and subjects are timed to see how long it takes them to read the standard passage. Suppose the analysis showed that subjects in the 'Times'-type (serif) font condition were able to read the passage in a significantly shorter period of time than subjects in the 'Arial'-type (sans serif) font condition.

The internal validity of this experiment would seem to be quite good. That is, one would probably be

on solid ground in concluding that the differences in time to read the passage were caused by the differences in the fonts. One might, however, have a number of questions about the external validity of this experiment. Some of these questions might include the following: Is the same effect found when the text is on paper with black on white usual-sized text? How about other sizes of text? How about other kinds of reading material? What about other populations than college students? How about elementary school children who are just learning to read? Would the same 'Times' font superiority effect be found in this population? How about older adults, etc.?

All these external validity questions concern the generalizability of the result to other contexts. It is worth noting that some of the characteristics of the experiment that make it have high internal validity (standardized text, conditions, etc.) are the very ones that limit its external validity.

*Construct validity.* Construct validity refers to the extent to which the interpretation of the constructs involved in an experiment is correct. Even if it is clear that the treatment actually caused the differences observed in the experiment (i.e., the experiment is internally valid), one might still question the interpretation of the constructs assumed to be measured by the independent or dependent variables.

A construct is a *theoretical* or *hypothetical* concept, variable, or idea that is presumed to underlie an *observed* variable or measurement operation. For example, an observed IQ test score is often assumed to measure the construct of 'intelligence.' The notion of a construct is also applicable to the manipulation of an independent variable as well.

For example, telling subjects that they are about to receive a painful shock might be assumed to manipulate the subjects' 'anxiety level'—*anxiety level* being the construct here. This was precisely the manipulation carried out in Schacter's (1959) well known experiment examining the relation between fear and affiliation. This experiment will serve to illustrate simply the issues in construct validity.

Schacter was interested in whether subjects who are experiencing more anxiety or fear have a stronger desire to affiliate with other people than subjects experiencing lower levels of fear. He carried out an experiment in which subjects were randomly assigned to be in either a 'high fear' condition or a 'low fear' condition. Subjects in the 'high fear' condition were told that the experiment they were in would require

them to receive intense, painful, but not permanently damaging electric shocks. Subjects in the 'low fear' condition were told that the experiment they were in would require them to receive mild and painless electric shocks. Both groups were told that the experimenter needed another ten minutes to prepare the equipment, but that the subject could wait either alone in a room with armchairs and magazines or in classroom with other subjects. The subject was asked to choose whether to wait alone or with others.

Schacter's hypothesis was that subjects in the 'high fear' condition would more often choose to wait with other subjects than those in the 'low fear' condition. The results were that the subjects in the high fear condition were almost twice as likely to choose to wait with other subjects than were subjects in the 'low fear' condition. His conclusion was that high fear increases the desire to affiliate with others.

The question of construct validity is applicable to both the independent and dependent variables here. For the independent variable, the construct validity question asks whether telling subjects they are about to receive an intense, painful shock actually manipulates the construct, 'fear,' rather than something else. For the dependent variable, the issue is whether a subject's choice to wait alone or with others actually is a measure of 'desire to affiliate with others' rather than something else.

Most researchers would probably agree that both the independent and dependent variables in Schacter's experiment appear to have good construct validity. In other experiments the answer does not always seem so clear. Maxwell and Delaney (1990) point out that if an experimenter wished to test the effect of 'death anxiety' on some dependent variable by showing some subjects a photograph of a dying person while other subjects were shown a neutral photo, another investigator might interpret any effects found as the effect of 'aroused compassion' on the dependent variable. It is not clear which construct is really being manipulated here.

*Statistical validity.* Statistical validity refers to the degree to which the statistical conclusions from the study are justified. The familiar Type I and Type II errors would represent threats to the statistical validity of a study. A Type I error occurs when an experimenter rejects a null hypothesis that is actually true. For example, if vitamin E actually has no effect on diastolic blood pressure, but the investigator in the vitamin E experiment by chance happens to select a sample in which a statistically significant effect is

found, a Type I error has been made. When there is no treatment effect (i.e., when the  $H_0$  is true) the probability of making a Type I error is, of course,  $\alpha$ , by convention usually chosen to be .05.

Because the Type I error rate is limited by the use of the conventional  $\alpha$  of .05, it is likely that Type I errors are not as much a problem as Type II errors. However, it is worth noting that in studies where a large number of statistical hypotheses are tested, it is virtually certain that at least some of the 'significant' results will be Type I errors. This is particularly problematic when the researcher mechanically carries out a large number of hypothesis tests (e.g., with a large correlation matrix) and then simply searches through the results to find significant values. The Type I error rate can be very high with this procedure.

A Type II error occurs when there is a real treatment effect in the population but the significance test fails to detect it (i.e., a nonsignificant result is obtained). The probability of making a Type II error when there is a real effect is often designated  $\beta$ . Consideration of Type II errors usually focuses not on  $\beta$ , but on  $1 - \beta$ , which is called the *power* of the test. The power of a hypothesis test is the probability that if there is a real effect it will be detected by the significance test (i.e., a significant result will be obtained).

Because conventional standards for the power of statistical tests are not as routinely applied as those for Type I errors, low power of statistical tests is a much more common problem than Type I errors. In evaluating an experimental design or a statistical result (particularly one which finds no significant effect), it is crucial to consider the power of the test. The power of a statistical test depends on a number of factors:

- $\alpha$  - the larger the  $\alpha$ , the higher the power
- Effect size* - the larger the effect size (difference between means), the higher the power
- $\sigma$  - the larger  $\sigma$  (within group standard deviation) is, the lower the power
- $n$  - the larger the sample size  $n$  is, the higher the power.

A simple example will illustrate the problem of low power. Suppose an investigator has developed a computer assisted instruction (CAI) program that she believes will improve mathematics performance of sixth graders. She designs an experiment to evaluate this hypothesis. She decides to randomly assign 20 sixth graders to each of three conditions (60 subjects total) and evaluate their mathematics performance

measured in grade equivalent units at the end of a two-week period. Grade equivalent units are a way of scaling performance so that the mean score for children at any particular grade level is the same as the grade level. For sixth graders, therefore, the average grade equivalent math score would be 6.

The three conditions she plans to use in her study are (1) a regular class control condition, (2) the two-week CAI condition, and (3) a computer use control condition in which children simply use a computer for other tasks during the time period the CAI group is using the program. In order to compute the power of the hypothesis test, reasonable estimates for values of all the factors mentioned above that influence power must be made. Suppose previous research has found that the standard deviation of grade equivalent scores for sixth graders is about  $\sigma = 2$ . Assume that the test will be made using  $\alpha = .05$ , and that the effect size she would be interested in detecting reflects the following population mean grade equivalent scores:

	<u>Regular Class</u>	<u>CAI</u>	<u>Computer Control</u>
Mean	6	7	6

Note that the effect size we would be interested in detecting here is a rather large one—one full grade level improvement after two weeks using the program. With  $n = 20$  per group and the above assumptions, the power of the  $F$ -test in a one way ANOVA can be found using standard formulas and tables. The *JMP* statistical package will also compute the power here very easily. It turns out to be .34. That is, even with a real effect as large as that specified and 20 subjects per group, the probability that her experiment will find a significant difference between the group means is only .34. Or, in other words, she has a 66% chance of finding no significant difference even if the program really has this very large effect. Of course, the chances of detecting a smaller effect are even less. This experiment is essentially a waste of time and effort. But note that the experimenter would not know this unless she had made the effort to do the power analysis.

How can the experiment be improved to increase the power of the statistical test? Of course, to the degree that any of the factors that influence power can be manipulated, the power will be affected. The most straightforward way to increase power is simply to increase the sample size  $n$ . What probability represents an adequate power? There is no unequivocal answer to this, but most researchers consider a power of at least .

80 to be minimally adequate in designing an experiment. The sample size  $n$  that would be required in this experiment to have a probability of .80 of detecting an effect the size we have specified can be computed (*JMP* and other packages will also do this). In this case the  $n$  turns out to be approximately 60 subjects per condition—considerably more than the 20 subjects per condition the experimenter initially planned.

The problem of low power in an experiment can lead to the waste of experimenter time and resources in carrying out a study in which the chances of detecting a real effect are low. But low power becomes truly devastating when one is attempting to make the claim that there is no substantial treatment effect. This kind of argument—sometimes referred to as trying to ‘prove the null hypothesis’—can arise in a number of ways. For example, perhaps two competing theories make different predictions about whether an effect should be found—one theory predicts an effect, the other predicts none. In order for the finding of no significant difference to provide support for the second theory, the experimenter must be able to make a convincing case that the power of the significance test was so high that even a small effect, if it were really there, would have been detected. In the absence of such a case, the argument falls flat. Why? Because it is the easiest thing in the world to carry out an experiment in which no significant differences are found. All one has to do is take a small enough sample size. The power will be so low that even a huge effect will very likely go undetected.

### Experimental Artifacts

Some of the major threats to the validity of experiments are referred to as experimental *artifacts*. An artifact is a source of confounded secondary variance that is the result of the artificial nature of the research process itself. Two independent variables are said to be *confounded* when they vary together in such a way that it is impossible to disentangle their effects on the dependent variable. That is, it is impossible to tell which variable is producing any effect found.

For example, suppose one were to do an experiment similar to vitamin E experiment but with just two groups—one group receiving the placebo, the other 200 IUs/day of vitamin E. If the experimenter were to decide, for some bizarre reason, to have only males in the placebo group and only females in the treatment group, the results of the experiment would

be uninterpretable. The reason is that the two variables, sex of subject and vitamin E dosage, are completely *confounded*—they vary together perfectly. If one were to find a large difference between the means for the conditions, it would be impossible to tell whether the difference resulted from the difference in dosage or the difference in sex of subjects.

This example also makes it clear that confounding is not always an all-or-none kind of problem. For example, if the experiment were carried out in such a way that, even though there were both males and females in each dosage condition, there was a substantial difference between the proportions of males and females in the two conditions, we would say that sex of subject was partially confounded with condition, and interpretation of any differences found would be problematic.

Confounding in an experiment may arise in a number of ways, but when it occurs as a result of the artificial nature of the research process it is referred to as an artifact. The phrase ‘artificial nature of the research process’ refers to the fact that most studies, as part of the process of measuring and manipulating variables, necessarily create at least a somewhat artificial situation. This artificiality can produce behavior in subjects that is different from what might ‘naturally’ occur.

### Hawthorne Effect

A classic example of an artifact is the famous ‘Hawthorne effect.’ A series of studies begun at the Hawthorne Western Electric plant in the 1920s investigated, among other things, how changes in physical conditions (lighting, crowding, etc.) in the plant affected worker productivity. Maxwell and Delaney (1990, p. 30-31) describe the problem that arose in interpreting the results.

When the brightness of the lights above a group of workers was increased, their performance improved. However, it was found that when the lighting for another selected group of workers was darkened somewhat, *their* performance also improved. In fact, it seemed that no matter what small change was made in the working environment of a group of workers, the result was an increase in their productivity. Although the investigators initially viewed the independent-variable construct merely as changes in level of illumination, that performance seemed to be affected similarly for the groups of workers being studied regardless of which feature of the physical

environment was manipulated led eventually to the conclusion that other constructs were being manipulated as well. The "Hawthorne effect" eventually came to be identified with the effect of psychological variables such as the perception of concern by management over working conditions or, more generally, the effects of awareness that one is participating in a research study.

#### *Response Set or Bias*

A number of problems related to subjects' response tendencies in research settings are referred to as artifacts under the general label of *response sets* or *biases*. A response set or response bias is any kind of stereotyped tendency to respond in a particular way regardless of the content of the item.

For example, an agreement bias would be a tendency towards always agreeing with statements the subject is asked to make a judgment about. In a series of items that ask a subject whether particular traits are descriptive of him, the subject might tend almost always to say "yes," regardless of the trait described in the item. Or, perhaps, when asked whether he agrees with a series of attitude statements, a subject might always have a bias to say that he disagreed with the statement no matter what its content. This would be a disagreement bias.

Agreement and disagreement biases can cause problems in the interpretation of measures that are composed of items worded in one direction only. If the items of a scale are worded so that higher scores on the scale are always associated with "yes" responses, then a high score on the scale could represent either a high level of the attribute being measured or, possibly, simply a high level of agreement bias. It is a good practice in constructing measures, therefore, to have half the items require subjects to respond in a positive manner, half in a negative manner to receive a high score on the scale. Individuals who are simply responding in accordance with an agreement or disagreement response bias, rather than to the actual content of the items, would then tend to have scores near the middle of the scale.

Another important kind of response bias is called a *social desirability* response bias. This bias is a tendency to respond always in a way that presents the self in as positive a light as possible. The observation in more recent times that some people tend to respond in only 'politically correct' way is a variation of this idea. Some subjects tend to respond in only socially acceptable ways, regardless of their true behavior,

attitudes, or beliefs. This tendency is sometimes called 'faking good,' and there are a number of scales that have been developed to attempt to measure the degree to which a subject is attempting to do this. Most large personality batteries contain scales designed to measure some variety of this tendency. They are sometimes called 'lie,' 'fake good,' or 'social desirability' scales. A well known scale of this type is the Marlowe-Crowne Social Desirability scale (Crowne & Marlowe, 1960).

It is difficult to eliminate totally problems of interpretation of scores when this kind of bias is present. Careful attention to wording of items so as to eliminate as much as possible social desirability cues is important. It is also possible to attempt to statistically adjust scores on a variable subject to this bias to try to eliminate its effects as long as a measure of social desirability is obtained along with the variable of interest. Another tack sometimes taken when a social desirability measure is available is to simply eliminate subjects who score high on social desirability from the main data analysis.

#### *Demand Characteristics*

A type of artifact that can lead to problems of interpretation, particularly in social psychological studies, results from the operation of what are called *demand characteristics*. It is important to recognize in designing and carrying out studies that as subjects participate in every stage of a study they typically try to identify exactly what it is that the experimenter expects of, or wants from, them. Subjects will use whatever cues are available that might inform them about what is expected or what their role should be. The cues in a research situation that influence subjects' perception of their role or what is expected of them are called the *demand characteristics* of the situation.

It is common for subjects to want to behave in an experiment in a way that is consistent with what is expected of them (though, of course, the opposite tendency may sometimes be present). This desire can lead subjects to behave in ways that are different from what their natural responses would be in that situation and thus, perhaps, to erroneous conclusions by the investigator. It is important, then, for investigators to pay careful attention to whatever demand characteristics might be present in the research situation that could affect subjects' behavior.

The fact that subjects' behavior is affected by what they believe is expected of them is one reason why it

is usually important that subjects not be made aware of what the investigator's experimental hypotheses are. In experiments where the hypotheses are known, or are obvious (e.g., in a drug study where informed consent requires subjects to know that several different drug conditions are involved), it is important that subjects be *blind* as to which condition they are actually in. This kind of study is called a *single blind* experiment. When not only subjects but also the experimenter who interacts with subjects is kept blind as to the experimental condition the subject is in, the experiment is said to be a *double-blind* experiment. Keeping the experimenter blind as to each subject's condition is intended to eliminate what are often called experimenter effects.

#### *Experimenter Effects*

Actually, any effect that an experimenter's expectancies might have on the outcome of an experiment is an instance of a larger class of artifacts referred to as *experimenter effects*. Experimenter effects refer to any kind of unintentional differential treatment of subjects in different conditions. This kind of differential treatment could lead to different results in the conditions that have nothing to do with the independent variable of interest. For example, a male experimenter might unintentionally, perhaps even unconsciously, treat male subjects differently than female subjects. Thus, sex of subject would be confounded with this unintentional differential treatment, and any sex effects found on the dependent variable would be uninterpretable.

Although experimenter effects might take a variety of forms, one commonly discussed problem is that of *experimenter expectancy bias*. The possibility that the experimenter's expectancies about the outcome of the experiment could affect subjects' behavior in such a way as to change the outcome was investigated extensively by Rosenthal (1976). Rosenthal and his colleagues have reported finding evidence of experimenter expectancy bias in a number of studies. Solso and Johnson (1994) describe the results of one such study:

An obvious demonstration of experimental bias was shown by Rosenthal and Fode (1963). A group of student experimenters who had some background in experimental psychology was asked to evaluate maze performance of two groups of rats. One group, so the experimenters were told, was selected from a long strain of "maze-bright" rats, while the second group was supposedly

selected from a long strain of "maze-dull" rats. The experimenters conducted a study of maze performance and, as expected, the maze-bright rats did significantly better than the maze-dull rats. The odd thing was that the rats were randomly selected from a standard sample of rats—there was no bright or dull distinction. Had the maze-bright rats actually performed better than the maze-dull rats? Probably not, but the experimenters who observed the bright rats expected them to perform better, and this expectation seemed to cloud their observations.

An additional possibility here, of course, is that the "maze-bright" rats actually did perform better than the other group because of some subtle difference in the kind of treatment they received from experimenters (perhaps kinder handling, or even just *more* handling).

It should be noted that some researchers (e.g., Barber, 1976) have criticized experimenter bias studies and have concluded that the strength of evidence suggesting the pervasiveness of such effects has been overstated.

## References

- Barber, T.X. (1976). *Pitfalls in human research: Ten pivotal points*. New York: Pergamon.
- Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.
- Kepler, J. (1984). *A defense of Tycho against Ursus*. In N. Jardine (trans. and ed.), *The birth of history and philosophy of science: Kepler's defense of Tycho against Ursus, with essays on its provenance and significance*. New York: Cambridge University Press. (Original work published 1601.)
- Maxwell, S.E. & Delaney, H.D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.
- Rosenthal, R. (1976). *Experimenter effects in behavioral research* (enlarged ed.) New York: Irvington.
- Rosenthal, R. & Fode, K. (1963). The effects of experimenter bias on the performance of the albino rat. *Behavioral Science, 8*, 183-189.
- Schacter, S. (1959). *The psychology of affiliation*. Stanford, CA: Stanford.
- Solso R.L. & Johnson, H.H. (1994). *Experimental psychology: A case approach* (Fifth ed.). New York: HarperCollins.