

This excerpt from

The Motion Aftereffect.
George Mather, Frans Verstraten and Stuart Anstis, editors.
© 1998 The MIT Press.

is provided in screen-viewable form for personal use only by members of MIT CogNet.

Unauthorized use or dissemination of this information is expressly forbidden.

If you have any questions about this material, please contact cognetadmin@cognet.mit.edu.

Chapter 6

Experimental Design in Psychological Research

Daniel J. Levitin

6.1 Introduction

Experimental design is a vast topic. As one thinks about the information derived from scientific studies, one confronts difficult issues in statistical theory and the limits of knowledge. In this chapter, we confine our discussion to a few of the most important issues in experimental design. This will enable students with no background in behavior research to critically evaluate psychological experiments, and to better understand the nature of empirical research in cognitive science.

Experimental psychology is a young science. The first laboratory of experimental psychology was established just over 100 years ago. Consequently, there are a great many mysteries about human behavior, perception, and performance that have not yet been solved. This makes it an exciting time to engage in psychological research—the field is young enough that there is still a great deal to do, and it is not difficult to think up interesting experiments. The goal of this chapter is to guide the reader in planning and implementing experiments, and in thinking about good experimental design.

A “good” experiment is one in which variables are carefully controlled or accounted for so that one can draw reasonable conclusions from the experiment’s outcome.

6.2 The Goals of Scientific Research

Generally, scientific research has four goals:

1. Description of behavior
2. Prediction of behavior
3. Determination of the causes of behavior
4. Explanations of behavior

These goals apply to the physical sciences as well as to the behavioral and life sciences. In basic science, the researcher’s primary concern is not with applications for a given finding. The goal of basic research is to increase our understanding of how the world works, or how things came to be the way they are.

Describing behavior impartially is the foremost task of the descriptive study, and because this is never completely possible, one tries to document any

From “Experimental Design in Psychoacoustic Research,” chapter 23 in *Music, Cognition, and Computerized Sound* (Cambridge, MA: MIT Press, 1999), 299–328. Reprinted with permission.

systematic biases that could influence descriptions (goal 1). By studying a phenomenon, one frequently develops the ability to *predict* certain behaviors or outcomes (goal 2), although prediction is possible without an understanding of underlying causes (we'll look at some examples in a moment). Controlled experiments are one tool that scientists use to reveal underlying causes so that they can advance from merely predicting behavior to understanding the *cause* of behavior (goal 3). *Explaining* behavior (goal 4) requires more than just a knowledge of causes; it requires a detailed understanding of the mechanisms by which the causal factors perform their functions.

To illustrate the distinction between the four goals of scientific research, consider the history of astronomy. The earliest astronomers were able to *describe* the positions and motions of the stars in the heavens, although they had no ability to *predict* where a given body would appear in the sky at a future date. Through careful observations and documentation, later astronomers became quite skillful at *predicting* planetary and stellar motion, although they lacked an understanding of the underlying factors that *caused* this motion. Newton's laws of motion and Einstein's special and general theories of relativity, taken together, showed that gravity and the contour of the space-time continuum cause the motions we observe. Precisely how gravity and the topology of space-time accomplish this still remains unclear. Thus, astronomy has advanced to the determination of causes of stellar motion (goal 3), although a full *explanation* remains elusive. That is, saying that gravity is responsible for astronomical motion only puts a name on things; it does not tell us how gravity actually works.

As an illustration from behavioral science, one might note that people who listen to loud music tend to lose their high-frequency hearing (description). Based on a number of observations, one can predict that individuals with normal hearing who listen to enough loud music will suffer hearing loss (prediction). A controlled experiment can determine that the loud music is the cause of the hearing loss (determining causality). Finally, study of the cochlea and basilar membrane, and observation of damage to the delicate hair cells after exposure to high-pressure sound waves, meets the fourth goal (explanation).

6.3 Three Types of Scientific Studies

In science there are three broad classes of studies: controlled studies, correlational studies, and descriptive studies. Often the type of study you will be able to do is determined by practicality, cost, or ethics, not directly by your own choice.

6.3.1 Controlled Studies ("True Experiments")

In a controlled experiment, the researcher starts with a group of subjects and randomly assigns them to an experimental condition. The point of *random assignment* is to control for extraneous variables that might affect the outcome of the experiment: variables that are different from the variable(s) being studied. With random assignment, one can be reasonably certain that any differences among the experimental groups were caused by the variable(s) manipulated in the experiment.

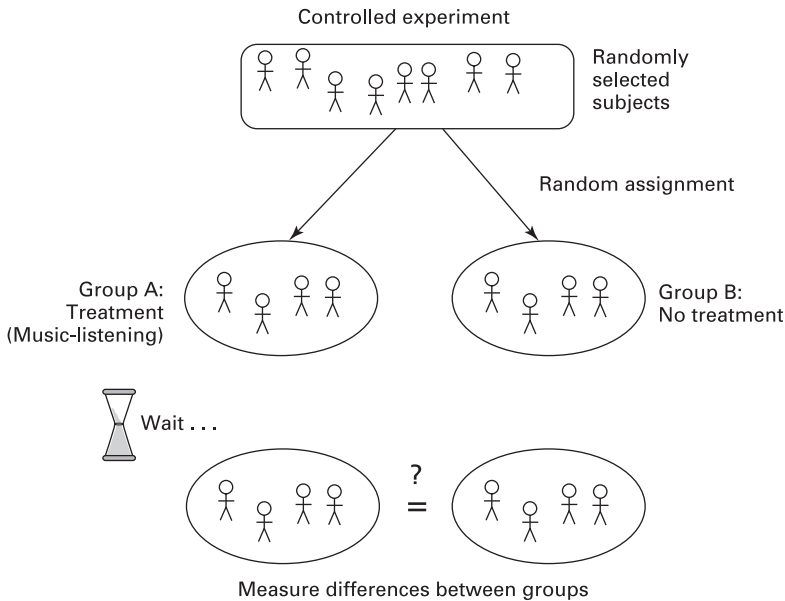


Figure 6.1
 In a controlled experiment, subjects are randomly assigned to conditions, and differences between groups are measured.

A controlled experiment in medical research might seek to discover if a certain food additive causes cancer. The researcher might randomly divide a group of laboratory mice into two smaller groups, giving the food additive to one group and not to the other. The variable he/she is interested in is the effect of the food additive; in the language of experimental design, this is called the “independent variable.” After a period of time, the researcher compares the mortality rates of the two groups; this quantity is called the “dependent variable” (figure 6.1). Suppose the group that received the additive tended to die earlier. In order to deduce that the additive caused the difference between the groups, the conditions must have been identical in every other respect. Both groups should have had the same diet, same feeding schedule, same temperature in their cages, and so on. Furthermore, the two groups of mice should have started out with similar characteristics, such as age, sex, and so on, so that these variables—being equally distributed between the two groups—can be ruled out as possible causes of the difference in mortality rates.

The two key components of a controlled experiment are *random assignment* of subjects, and *identical experimental conditions* (see figure 6.1). A researcher might have a hypothesis that people who study for an exam while listening to music will score better than people who study in silence. In the language of experimental design, music-listening is the *independent variable*, and test performance, the quantity to be measured, is the *dependent variable*.

No one would take this study seriously if the subjects were divided into two groups based on how they did on the previous exam—if, for instance, the top half of the students were placed in the music-listening condition, and the

bottom half of the students in the silence condition. Then if the result of the experiment was that the music listeners as a group tended to perform better on their next exam, one could argue that this was not because they listened to music while they studied, but because they were the better students to begin with.

Again, the theory behind random assignment is to have groups of subjects who start out the same. Ideally, each group will have similar distributions on every conceivable dimension—age, sex, ethnicity, IQ, and variables that you might not think are important, such as handedness, astrological sign, or favorite television show. Random assignment makes it unlikely that there will be any large systematic differences between the groups.

A similar design flaw would arise if the *experimental conditions* were different. For example, if the music-listening group studied in a well-lit room with windows, and the silence group studied in a dark, windowless basement, any difference between the groups could be due to the different environments. The room conditions become confounded with the music-listening conditions, such that it is impossible to deduce which of the two is the causal factor.

Performing random assignment of subjects is straightforward. Conceptually, one wants to mix the subjects' names or numbers thoroughly, then draw them out of a hat. Realistically, one of the easiest ways to do this is to generate a different random number for each subject, and then sort the random numbers. If n equals the total number of subjects you have, and g equals the number of groups you are dividing them into, the first n/g subjects will comprise the first group, the next n/g will comprise the second group, and so on.

If the results of a controlled experiment indicate a difference between groups, the next question is whether these findings are generalizable. If your initial group of subjects (the large group, before you randomly assigned subjects to conditions) was also randomly selected (called *random sampling* or *random selection*, as opposed to *random assignment*), this is a reasonable conclusion to draw. However, there are almost always some constraints on one's initial choice of subjects, and this constrains generalizability. For example, if all the subjects you studied in your music-listening experiment lived in fraternities, the finding might not generalize to people who do not live in fraternities. If you want to be able to generalize to all college students, you would need to take a representative sample of all college students. One way to do this is to choose your subjects randomly, such that each member of the population you are considering (college students) has an equal likelihood of being placed in the experiment.

There are some interesting issues in representative sampling that are beyond the scope of this chapter. For example, if you wanted to take a representative sample of all American college students and you chose American college students randomly, it is possible that you would be choosing several students from some of the larger colleges, such as the University of Michigan, and you might not choose any students at all from some of the smaller colleges, such as Bennington College; this would limit the applicability of your findings to the colleges that were represented in your sample. One solution is to conduct a *stratified sample*, in which you first randomly select colleges (making it just as likely that you'll choose large and small colleges) and then randomly select the

same number of students from each of those colleges. This ensures that colleges of different sizes are represented in the sample. You then weight the data from each college in accordance with the percentage contribution each college makes to the total student population of your sample. (For further reading, see Shaughnessy and Zechmeister 1994.)

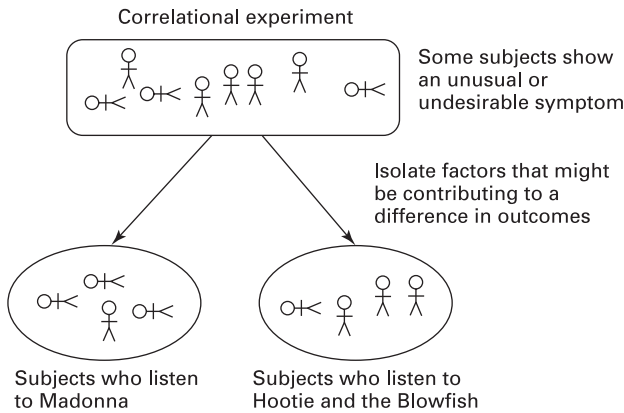
Choosing subjects randomly requires careful planning. If you try to take a random sample of Stanford students by standing in front of the Braun Music Building and stopping every third person coming out, you might be selecting a greater percentage of music students than actually exists on campus. Yet truly random samples are not always practical. Much psychological research is conducted on college students who are taking an introductory psychology class, and are required to participate in an experiment for course credit. It is not at all clear whether American college students taking introductory psychology are representative of students in general, or of people in the world in general, so one should be careful not to overgeneralize findings from these studies.

6.3.2 Correlational Studies

A second type of study is the *correlational study* (figure 6.2). Because it is not always practical or ethical to perform random assignments, scientists are sometimes forced to rely on patterns of co-occurrence, or correlations between events. The classic example of a correlational study is the link between cigarette smoking and cancer. Few educated people today doubt that smokers are more likely to die of lung cancer than are nonsmokers. However, in the history of scientific research there has never been a controlled experiment with human subjects on this topic. Such an experiment would take a group of healthy nonsmokers, and randomly assign them to two groups, a smoking group and a nonsmoking group. Then the experimenter would simply wait until most of the people in the study have died, and compare the average ages and causes of death of the two groups. Because our hypothesis is that smoking causes cancer, it would clearly be unethical to ask people to smoke who otherwise would not.

The scientific evidence we have that smoking causes cancer is correlational. That is, when we look at smokers as a group, a higher percentage of them do indeed develop fatal cancers, and die earlier, than do nonsmokers. But without a controlled study, the possibility exists that there is a third factor—a mysterious “factor x”—that both causes people to smoke and to develop cancer. Perhaps there is some enzyme in the body that gives people a nicotine craving, and this same enzyme causes fatal cancers. This would account for both outcomes, the kinds of people who smoke and the rate of cancers among them, and it would show that there is no causal link between smoking and cancer.

In correlational studies, a great deal of effort is devoted to trying to uncover differences between the two groups studied in order to identify any causal factors that might exist. In the case of smoking, none have been discovered so far, but the failure to discover a third causal factor does not prove that one does not exist. It is an axiom in the philosophy of science that one can prove only the presence of something; one can't prove the absence of something—it could always be just around the corner, waiting to be discovered in the next experiment (Hempel 1966). In the real world, behaviors and diseases are usually brought



Two possible conclusions:

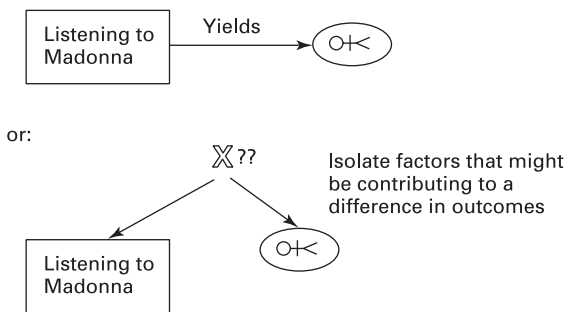


Figure 6.2 In a correlational study, the researcher looks for a relation between two observed behaviors—in this case, the relation between untimely death and listening to Madonna recordings.

on by a number of complicated factors, so the mysterious third variable, “factor x,” could in fact be a collection of different, and perhaps unrelated, variables that act together to cause the outcomes we observe.

An example of a correlational study with a hypothesized musical cause is depicted in figure 6.2. Such a study would require extensive interviews with the subjects (or their survivors), to try to determine all factors that might separate the subjects exhibiting the symptom from the subjects without the symptom.

The problem with correlational studies is that the search for underlying factors that account for the differences between groups can be very difficult. Yet many times, correlational studies are all we have, because ethical considerations preclude the use of controlled experiments.

6.3.3 Descriptive Studies

Descriptive studies do not look for differences between people or groups, but seek only to describe an aspect of the world as it is. A descriptive study in physics might seek to discover what elements make up the core of the planet Jupiter. The goal in such a study would not be to compare Jupiter’s core with

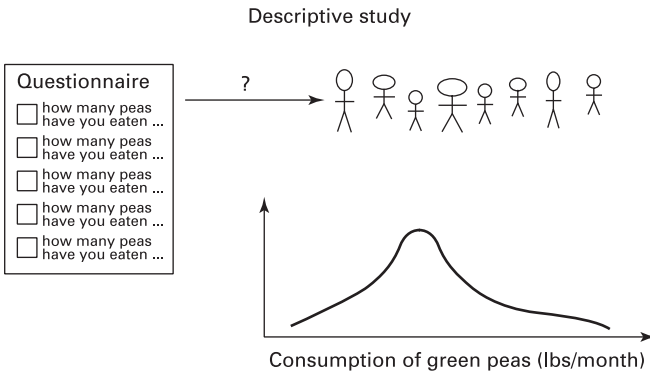


Figure 6.3

In a descriptive study, the researcher seeks to describe some aspect of the state of the world, such as people's consumption of green peas.

the core of other planets, but to learn more about the origins of the universe. In psychology, we might want to know the part of the brain that is activated when someone performs a mental calculation, or the number of pounds of fresh green peas the average Canadian eats in a year (figure 6.3). Our goal in these cases is not to contrast individuals but to acquire some basic data about the nature of things. Of course, descriptive studies can be used to establish “norms,” so that we can compare people against the average, but as their name implies, the primary goal in descriptive experiments is often just to describe something that had not been described before. Descriptive studies are every bit as useful as controlled experiments and correlational studies—sometimes, in fact, they are even more valuable because they lay the foundation for further experimental work.

6.4 Design Flaws in Experimental Design

6.4.1 Clever Hans

There are many examples of flawed studies or flawed conclusions that illustrate the difficulties in controlling extraneous variables. Perhaps the most famous case is that of Clever Hans.

Clever Hans was a horse owned by a German mathematics teacher around the turn of the twentieth century. Hans became famous following many demonstrations in which he could perform simple addition and subtraction, read German, and answer simple questions by tapping his hoof on the ground (Watson 1967). One of the first things that skeptics wondered (as you might) is whether Hans would continue to be clever when someone other than his owner asked the questions, or when Hans was asked questions that he had never heard before. In both these cases, Hans continued to perform brilliantly, tapping out the sums or differences for arithmetic problems.

In 1904, a scientific commission was formed to investigate Hans's abilities more carefully. The commission discovered, after rigorous testing, that Hans could never answer a question if the questioner did not also know the answer,

or if Hans could not see his questioner. It was finally discovered that Hans had become very adept at picking up subtle (and probably unintentional) movements on the part of the questioner that cued him as to when he should stop tapping his foot. Suppose a questioner asked Hans to add 7 and 3. Hans would start tapping his hoof, and keep on tapping until the questioner stopped him by saying “Right! Ten!” or, more subtly, by moving slightly when the correct answer was reached.

You can see how important it is to ensure that extraneous cues or biases do not intrude into an experimental situation.

6.4.2 *Infants’ Perception of Musical Structure*

In studies of infants’ perception of music, infants typically sit in their mother’s lap while music phrases are played over a speaker. Infants tend to turn their heads toward a novel or surprising event, and this is the dependent variable in many infant studies; the point at which the infants turn their heads indicates when they perceive a difference in whatever is being played. Suppose you ran such a study and found that the infants were able to distinguish Mozart selections that were played normally from selections of equal length that began or ended in the middle of a musical phrase. You might take this as evidence that the infants have an innate understanding of musical phraseology.

Are there alternative explanations for the results? Suppose that in the experimental design, the mothers could hear the music, too. The mothers might unconsciously cue the infants to changes in the stimulus that they (the mothers) detect. A simple solution is to have the mothers wear headphones playing white noise, so that their perception of the music is masked.

6.4.3 *Computers, Timing, and Other Pitfalls*

It is very important that you not take anything for granted as you design a careful experiment, and control extraneous variables. For example, psychologists studying visual perception frequently present their stimuli on a computer using the MacIntosh or Windows operating system. In a computer program, the code may specify that an image is to remain on the computer monitor for a precise number of milliseconds. Just because you specify this does not make it happen, however. Monitors have a *refresh rate* (60 or 75 Hz is typical), so the “on time” of an image will always be an integer multiple of the refresh cycle (13.33 milliseconds for a 75 Hz refresh rate) no matter what you instruct the computer to do in your code. To make things worse, the MacIntosh and Windows operating systems do not guarantee “refresh cycle accuracy” in their updating, so an instruction to put a new image on the screen may be delayed an unknown amount of time.

It is important, therefore, always to verify, using some external means, that the things you think are happening in your experiment are actually happening. Just because you leave the volume control on your amplifier at the same spot doesn’t mean the volume of a sound stimulus you are playing will be the same from day to day. You should measure the output and not take the knob position for granted. Just because a frequency generator is set for 1000 Hz does not mean it is putting out a 1000 Hz signal. It is good science for you to measure the output frequency yourself.

6.5 Number of Subjects

How many subjects are enough? In statistics, the word “population” refers to the total group of people to which the researcher wishes to generalize findings. The population might be female sophomores at Stanford, or all Stanford students, or all college students in the United States, or all people in the United States. If one is able to draw a representative sample of sufficient size from a population, one can make inferences about the whole population based on a relatively small number of cases. This is the basis of presidential polls, for example, in which only 2000 voters are surveyed, and the outcome of an election can be predicted with reasonable accuracy.

The size of the sample required is dependent on the degree of homogeneity or heterogeneity in the total population you are studying. In the extreme, if you are studying a population that is so homogeneous that every individual is identical on the dimensions being studied, a sample size of one will provide all the information you need. At the other extreme, if you are studying a population that is so heterogeneous that each individual differs categorically on the dimension you are studying, you will need to sample the entire population.

As a “rough-and-ready” rule, if you are performing a descriptive perceptual experiment, and the phenomenon you are studying is something that you expect to be invariant across people, you need to use only a few subjects, perhaps five. An example of this type of study would be calculating threshold sensitivities for various sound frequencies, such as was done by Fletcher and Munson (1933).

If you are studying a phenomenon for which you expect to find large individual differences, you might need between 30 and 100 subjects. This depends to some degree on how many different conditions there are in the study. In order to obtain means with a relatively small variance, it is a good idea to have at least five to ten subjects in each experimental condition.

6.6 Types of Experimental Designs

Suppose you are researching the effect of music-listening on studying efficiency, as mentioned at the beginning of this chapter. Let’s expand on the simpler design described earlier. You might divide your subjects into five groups: two experimental groups and three control groups. One experimental group would listen to rock music, and the other would listen to classical music. Of the three control groups, one would listen to rock music for the same number of minutes per day as the experimental group listening to rock (but not while they were studying); a second would do the same for classical music; the third would listen to no music at all. This is called a *between-subjects* design, because each subject is in one condition and one condition only (also referred to as an *independent groups* design). If you assign 10 subjects to each experimental condition, this would require a total of 50 subjects. Table 6.1 shows the layout of this experiment. Each distinct box in the table is called a *cell* of the experiment, and subject numbers are filled in for each cell. Notice the asymmetry for the *no music* condition. The experiment was designed so that there is only one “no music” condition, whereas there are four music conditions of various types.

Table 6.1
Between-subjects experiment on music and study habits

Condition	Only while studying	Only while not studying
<i>Music</i>		
Classical	Subjects 1–10	Subjects 11–20
Rock	Subjects 21–30	Subjects 31–40
<i>No music</i>		
	Subjects 41–50	Subjects 41–50

Testing 50 subjects might not be practical. An alternative is a *within-subjects* design, in which every subject is tested in every condition (also called a *repeated measures* design). In this example, a total of ten subjects could be randomly divided into the five conditions, so that two subjects experience each condition for a given period of time. Then the subjects switch to another condition. By the time the experiment is completed, ten observations have been collected in each cell, and only ten subjects are required.

The advantage of each subject experiencing each condition is that you can obtain measures of how each individual is affected by the manipulation, something you cannot do in the between-subjects design. It might be the case that some people do well in one type of condition and other people do poorly in it, and the within-subjects design is the best way to show this. The obvious advantage to the within-subjects design is the smaller number of subjects required. But there are disadvantages as well.

One disadvantage is *demand characteristics*. Because each subject experiences each condition, they are not as naive about the experimental manipulation. Their performance could be influenced by a conscious or unconscious desire to make one of the conditions work better. Another problem is *carryover effects*. Suppose you were studying the effect of Prozac on learning, and that the half-life of the drug is 48 hours. The group that gets the drug first might still be under its influence when they are switched to the nondrug condition. This is a carryover effect. In the music-listening experiment, it is possible that listening to rock music creates anxiety or exhilaration that might last into the next condition.

A third disadvantage of within-subjects designs is *order effects*, and these are particularly troublesome in psychophysical experiments. An order effect is similar to a carryover effect, and it concerns how responses in an experiment might be influenced by the order in which the stimuli or conditions are presented. For instance, in studies of speech discrimination, subjects can habituate (become used to, or become more sensitive) to certain sounds, altering their threshold for the discriminability of related sounds. A subject who habituates to a certain sound may respond differently to the sound immediately following it than he/she normally would. For these reasons, it is important to counterbalance the order of presentations; presenting the same order to every subject makes it difficult to account for any effects that are due merely to order.

One way to reduce order effects is to present the stimuli or conditions in random order. In some studies, this is sufficient, but to be really careful about order effects, the random order simply is not rigorous enough. The solution is to use every possible order. In a *within-subjects* design, each subject would

complete the experiment with each order. In a *between-subjects* design, different subjects would be assigned different orders. The choice will often depend on the available resources (time and availability of subjects). The number of possible orders is $N!$ ("n factorial"), where N equals the number of stimuli. With two stimuli there are two possible orders ($2! = 2 \times 1$); with three stimuli there are six possible orders ($3! = 3 \times 2 \times 1$); with six stimuli there are 720 possible orders ($6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$). Seven hundred twenty orders is not practical for a within-subjects design, or for a between-subjects design. One solution in this case is to create an order that presents each stimulus in each serial position. A method for accomplishing this involves using the Latin Square. For even-numbered N , the size of the Latin Square will be $N \times N$; therefore, with six stimuli you would need only 36 orders, not 720. For odd-numbered N , the size of the Latin Square will be $N \times 2N$. Details of this technique are covered in experimental design texts such as Kirk (1982) and Shaughnessy and Zechmeister (1994).

6.7 Ethical Considerations in Using Human Subjects

Some experiments on human subjects in the 1960s and 1970s raised questions about how human subjects are treated in behavioral experiments. As a result, guidelines for human experimentation were established. The American Psychological Association, a voluntary organization of psychologists, formulated a code of ethical principles (American Psychological Association 1992). In addition, most universities have established committees to review and approve research using human subjects. The purpose of these committees is to ensure that subjects are treated ethically, and that fair and humane procedures are followed. In some universities, experiments performed for course work or experiments done as "pilot studies" do not require approval, but these rules vary from place to place, so it is important to determine the requirements at your institution before engaging in any human subject research.

It is also important to understand the following four basic principles of ethics in human subject research:

1. *Informed consent.* Before agreeing to participate in an experiment, subjects should be given an accurate description of their task in the experiment, and told any risks involved. Subjects should be allowed to decline, or to discontinue participation in the experiment at any time without penalty.
2. *Debriefing.* Following the experiment, the subjects should be given an explanation of the hypothesis being tested and the methods used. The experimenter should answer any questions the subjects have about the procedure or hypothesis. Many psychoacoustic experiments involve difficult tasks, leading some subjects to feel frustrated or embarrassed. Subjects should never leave an experiment feeling slow, stupid, or untalented. It is the experimenter's responsibility to ensure that the subjects understand that these tasks are inherently difficult, and when appropriate, the subjects should be told that the data are not being used to evaluate them personally, but to collect information on how the population in general can perform the task.

3. *Privacy and confidentiality.* The experimenter must carefully guard the data that are collected and, whenever possible, code and store the data in such a way that subjects' identities remain confidential.

4. *Fraud.* This principle is not specific to human subjects research, but applies to all research. An essential ethical standard of the scientific community is that scientific researchers never fabricate data, and never knowingly, intentionally, or through carelessness allow false data, analyses, or conclusions to be published. Fraudulent reporting is one of the most serious ethical breaches in the scientific community.

6.8 Analyzing Your Data

6.8.1 Quantitative Analysis

Measurement Error Whenever you measure a quantity, there are two components that contribute to the number you end up with: the actual value of the thing you are measuring and some amount of measurement error, both human and mechanical. It is an axiom of statistics that measurement error is just as likely to result in an overestimate as an underestimate of the true value. That is, each time you take a measurement, the error term (let's call it *epsilon*) is just as likely to be positive as negative. Over a large number of measurements, the positive errors and negative errors will cancel out, and the average value of epsilon will approach 0. The larger the number of measurements you make, the closer you will get to the true value. Thus, as the number of measurements approaches infinity, the arithmetic average of your measurements approaches the true quantity being measured. Suppose we are measuring the weight of a sandbag.

Formally, we would write:

$$n \rightarrow \infty, \quad \bar{\epsilon} = 0$$

where $\bar{\epsilon}$ = the mean of epsilon, and

$$n \rightarrow \infty, \quad \bar{w} = w$$

where \bar{w} = the mean of all the weight measurements and w = the true weight.

When measuring the behavior of human subjects on a task, you encounter not only measurement error but also performance error. The subjects will not perform identically every time. As with measurement error, the more observations you make, the more likely it is that the performance errors cancel each other out. In psychoacoustic tasks the performance errors can often be relatively large. This is the reason why one usually wants to have the subject perform the same task many times, or to have many subjects perform the task a few times.

Because of these errors, the value of your dependent variable(s) at the end of the experiment will always deviate from the true value by some amount. Statistical analysis helps in interpreting these differences (Bayesian inferencing, meta-analyses, effect size analysis, significance testing) and in predicting the true value (point estimates and confidence intervals). The mechanics of these

tests are beyond the scope of this chapter, and the reader is referred to the statistics textbooks mentioned earlier.

Significance Testing Suppose you wish to observe differences in interval identification ability between brass players and string players. The question is whether the difference you observe between the two groups can be wholly accounted for by measurement and performance error, or whether a difference of the size you observe indicates a true difference in the abilities of these musicians.

Significance tests provide the user with a “p value,” the probability that the experimental result could have arisen by chance. By convention, if the p value is less than .05, meaning that the result could have arisen by chance less than 5% of the time, scientists accept the result as statistically significant. Of course, $p < .05$ is arbitrary, and it doesn’t deal directly with the opposite case, the probability that the data you collected indicate a genuine effect, but the statistical test failed to detect it (a power analysis is required for this). In many studies, the probability of failing to detect an effect, when it exists, can soar to 80% (Schmidt 1996). An additional problem with a criterion of 5% is that a researcher who measures 20 different effects is likely to measure one as significant by chance, even if no significant effect actually exists.

Statistical significance tests, such as the analysis of variance (ANOVA), the f-test, chi-square test, and t-test, are methods to determine the probability that observed values in an experiment differ only as a result of measurement errors. For details about how to choose and conduct the appropriate tests, or to learn more about the theory behind them, consult a statistics textbook (e.g., Daniel 1990; Glenberg 1988; Hayes 1988).

Alternatives to Classical Significance Testing Because of problems with traditional significance testing, there is a movement, at the vanguard of applied statistics and psychology, to move away from “p value” tests and to rely on alternative methods, such as Bayesian inferencing, effect sizes, confidence intervals, and meta-analyses (refer to Cohen 1994; Hunter and Schmidt 1990; Schmidt 1996). Yet many people persist in clinging to the belief that the most important thing to do with experimental data is to test them for statistical significance. There is great pressure from peer-reviewed journals to perform significance tests, because so many people were taught to use them. The fact is, the whole point of significance testing is to determine whether a result is repeatable when one doesn’t have the resources to repeat an experiment.

Let us return to the hypothetical example mentioned earlier, in which we examined the effect of music on study habits using a “within-subjects” design (each subject is in each condition). One possible outcome is that the difference in the mean test scores among groups was not significantly different by an analysis of variance (ANOVA). Yet suppose that, ignoring the means, every subject in the music-listening condition had a higher score than in the no-music condition. We are not interested in the size of the difference now, only in the direction of the difference. The null hypothesis predicts that the manipulation would have no effect at all, and that half of the subjects should show a difference in one direction and half in the other. The probability of all 10 subjects showing an effect in the same direction is $1/2^{10}$ or 0.0009, which is highly

significant. Ten out of 10 subjects indicates *repeatability*. The technique just described is called the *sign test*, because we are looking only at the arithmetic sign of the differences between groups (positive or negative).

Often, a good alternative to significance tests is estimates of *confidence intervals*. These determine with a given probability (e.g., 95%) the range of values within which the true population parameters lie. Another alternative is an analysis of *conditional probabilities*. That is, if you observe a difference between two groups on some measure, determine whether a subject's membership in one group or the other will improve your ability to predict his/her score on the dependent variable, compared with not knowing what group he/she was in (an example of this analysis is in Levitin 1994a). A good overview of these alternative statistical methods is contained in the paper by Schmidt (1996).

Aside from statistical analyses, in most studies you will want to compute the mean and standard deviation of your dependent variable. If you had distinct treatment groups, you will want to know the individual means and standard deviations for each group. If you had two continuous variables, you will probably want to compute the *correlation*, which is an index of how much one variable is related to the other. Always provide a table of means and standard deviations as part of your report.

6.8.2 *Qualitative Analysis, or "How to Succeed in Statistics without Significance Testing"*

If you have not had a course in statistics, you are probably at some advantage over anyone who has. Many people who have taken statistics courses rush to plug the numbers into a computer package to test for statistical significance. Unfortunately, students are not always perfectly clear on exactly what it is they are testing or why they are testing it.

The first thing one should do with experimental data is to graph them in a way that clarifies the relation between the data and the hypothesis. Forget about statistical significance testing—what does the pattern of data suggest? Graph everything you can think of—individual subject data, subject averages, averages across conditions—and see what patterns emerge. Roger Shepard has pointed out that the human brain is not very adept at scanning a table of numbers and picking out patterns, but is much better at picking out patterns in a visual display.

Depending on what you are studying, you might want to use a bar graph, a line graph, or a bivariate scatter plot. As a general rule, even though many of the popular graphing and spreadsheet packages will allow you to make pseudo-three-dimensional graphs, don't ever use three dimensions unless the third dimension actually represents a variable. Nothing is more confusing than a graph with extraneous information.

If you are making several graphs of the same data (such as individual subject graphs), make sure that each graph is the same size and that the axes are scaled identically from one graph to another, in order to facilitate comparison. Be sure all your axes are clearly labeled, and don't divide the axis numbers into units that aren't meaningful (for example, in a histogram with "number of subjects" on the ordinate, the scale shouldn't include half numbers because subjects come only in whole numbers).

Use a line graph if your variables are continuous. The lines connecting your plot points imply a continuous variable. Use a bar graph if the variables are categorical, so that you don't fool the reader into thinking that your observations were continuous. Use a bivariate scatter plot when you have two continuous variables, and you want to see how a change in one variable affects the other variable (such as how IQ and income might correlate). Do *not* use a bivariate scatterplot for categorical data. (For more information on good graph design, see Chambers et al. 1983; Cleveland 1994; Kosslyn 1994).

Once you have made all your graphs, look them over for interesting patterns and effects. Try to get a feel for what you have found, and understand how the data relate to your hypotheses and your experimental design. A well-formed graph can make a finding easy to understand and evaluate far better than a dry recitation of numbers and statistical tests can do.

Acknowledgments

This chapter benefited greatly from comments by Perry Cook, Lynn Gerow, Lewis R. Goldberg, John M. Kelley, and John Pierce. During the preparation of this chapter, I received direct support from an ONR graduate research fellowship (N-00014-89-J-3186), and indirect support from CCRMA and from an ONR Grant to M. I. Posner (N-00014-89-3013).

References

- American Psychological Association. (1992). "Ethical Principles of Psychologists and Code of Conduct." *American Psychologist*, 47, 1597–1611.
- American Psychological Association. (1994). *Publication Manual of the American Psychological Association*. Fourth edition. Washington, D.C.: American Psychological Association.
- Butler, D., and W. D. Ward. (1988). "Effacing the Memory of Musical Pitch." *Music Perception*, 5 (3), 251–260.
- Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. (1983). *Graphical Methods for Data Analysis*. New York: Chapman & Hall.
- Cleveland, W. S. (1994). *The Elements of Graphing Data*. Revised edition. Summit, N.J.: Hobart Press.
- Cohen, J. (1994). "The Earth Is Round ($p < .05$)." *American Psychologist*, 49, 997–1003.
- Cozby, P. C. (1989). *Methods in Behavioral Research*. Fourth edition. Mountain View, Calif.: Mayfield Publishing Co.
- Daniel, W. W. (1990). *Applied Nonparametric Statistics*. Second edition. Boston: PWS-Kent.
- Deutsch, D. (1991). "The Tritone Paradox: An Influence of Language on Music Perception." *Music Perception*, 84, 335–347.
- Deutsch, D. (1992). "The Tritone Paradox: Implications for the Representation and Communication of Pitch Structure." In M. R. Jones and S. Holleran, eds., *Cognitive Bases of Musical Communication*. Washington, D.C.: American Psychological Association.
- Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.
- Fletcher, H., and W. A. Munson. (1933). "Loudness, Its Definition, Measurement and Calculation." *Journal of the Acoustical Society of America*, 72, 82–108.
- Glenberg, A. (1988). *Learning from Data: An Introduction to Statistical Reasoning*. San Diego: Harcourt, Brace, Jovanovich.
- Hayes, W. (1988). *Statistics*. Fourth edition. New York: Holt, Rinehart and Winston.
- Hempel, C. G. (1966). *Philosophy of Natural Science*. Englewood Cliffs, N.J.: Prentice-Hall.
- Hunter, J. E., and F. L. Schmidt. (1990). *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Newbury Park, Calif.: Sage.
- Kirk, R. E. (1982). *Experimental Design: Procedures for the Behavioral Sciences*. Second edition. Pacific Grove, Calif.: Brooks/Cole.
- Kosslyn, S. M. (1994). *Elements of Graph Design*. New York: Freeman.
- Levitin, D. J. (1994a). "Absolute Memory for Musical Pitch: Evidence from the Production of Learned Melodies." *Perception & Psychophysics*, 56 (4), 414–423.

- . (1994b). *Problems in Applying the Kolmogorov-Smirnov Test: The Need for Circular Statistics in Psychology*. Technical Report #94-07. University of Oregon, Institute of Cognitive & Decision Sciences.
- Schmidt, F. L. (1996). "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers." *Psychological Methods*, VI (2): 115–129.
- Shaughnessy, J. J., and E. B. Zechmeister. (1994). *Research Methods in Psychology*. Third edition. New York: McGraw-Hill.
- Stern, A. W. (1993). "Natural Pitch and the A440 Scale." Stanford University, CCRMA. (Unpublished report).
- Watson, J. B. (1967). *Behavior: An Introduction to Comparative Psychology*. New York: Holt, Rinehart and Winston. First published 1914.
- Zar, J. H. (1984). *Biostatistical Analysis*. Second edition. Englewood Cliffs, N.J.: Prentice-Hall.

This excerpt from

The Motion Aftereffect.

George Mather, Frans Verstraten and Stuart Anstis, editors.

© 1998 The MIT Press.

is provided in screen-viewable form for personal use only by members of MIT CogNet.

Unauthorized use or dissemination of this information is expressly forbidden.

If you have any questions about this material, please contact cognetadmin@cognet.mit.edu.