

# Experimental Design in Psychoacoustic Research

by Daniel J. Levitin

From the book

"Music, Cognition and Computerized Sound : An Introduction to Psychoacoustics."

P. R. Cook (Editor), M.I.T. Press, 1999.

---

## Chapter 23: Experimental Design in Psychoacoustic Research

By Daniel J. Levitin

### 23.1 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 the space available for this chapter, we confine our discussion to a few of the most important issues in experimental design. This will enable students with no prior background in behavior research to complete a simple research project. The student will need to consult statistical and experimental design texts (as listed in the references at the end of this chapter) in order to answer specific questions that are not covered here.

This chapter is intended for undergraduate and beginning graduate level students enrolled in a psychoacoustics course, or who are interested in learning about psychoacoustics research through independent study. The chapter assumes no prior knowledge of statistics, probability, experimental design, or psychology. A central component of the psychoacoustics course as it has been taught at Stanford for the past seven years is the term project. Typically this is in the form of a psychoacoustic experiment, using human subjects to test a hypothesis or demonstrate a principle in psychoacoustics or auditory perception. The term project affords students the opportunity to participate first hand in psychoacoustics research, and it encourages them to become engaged with a rich and interesting history of behavioral research.

Experimental psychology is a young science. The first laboratory of experimental psychology was established just over one hundred years ago. As such, there are a great many mysteries about human behavior, perception, and performance that have not yet been solved. This makes it an exciting time in history 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.

## 23.2 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, and
- (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 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 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 motion 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; still 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. An explanation for 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 an *explanation* remains elusive at this time (Goal 4). That is, saying that gravity is responsible for astronomical motion only puts a name on things, it doesn't tell us how gravity 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 as well (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 the basilar membrane, and observation of damage to the delicate hair cells after exposure to high pressure sound waves meets the fourth goal (explanation).

## 23.3 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.

### 23.3.1 Controlled studies (or "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 than the variable(s) he is studying. With random assignment, one can be reasonably certain that any differences among the experimental groups were caused by the variable(s) that were manipulated in the experiment.

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 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 would compare the mortality rates of the two groups; this quantity is called the "dependent variable." Suppose that the group that received the additive tended to die earlier. In order to correctly 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 schedules, 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.

Of course, subjects (both humans and mice) differ from one another in more ways than we can ever document. The ideal experiment would use truly identical subjects in each condition, but because this is impossible, the technique of *random assignment* is used. We assume that whatever differences existed among individual subjects prior to the intervention are just as likely to be found in one group as the other.

The two key components of a controlled experiment are *random assignment* of subjects, and *identical experimental conditions* (see Figure 23.1). A researcher might have a hypothesis that people who study for an exam while listening to music score better on the exam 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 other variables that you might not think are important, such as handedness, astrological sign, or favorite

television show. With random assignment, we make it unlikely for there to 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.

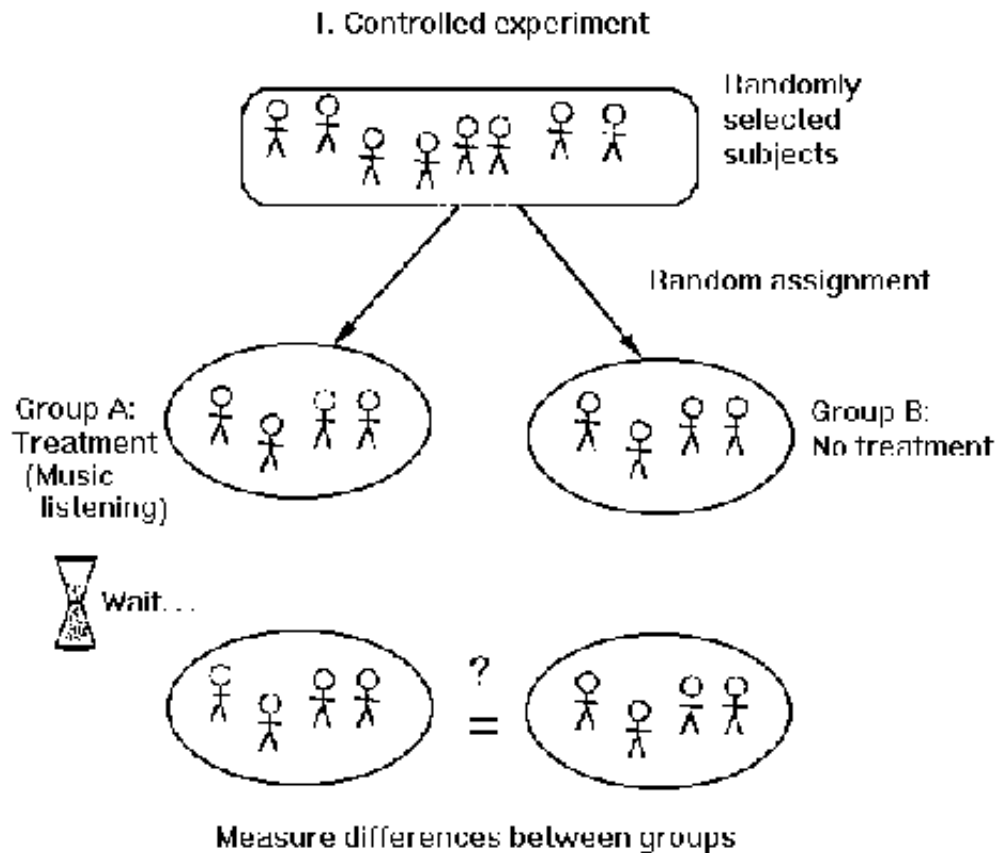


Figure 23.1 In a controlled experiment, subjects are randomly assigned to conditions, and differences between groups are measured.

Performing random assignment of subjects is straightforward. Conceptually, one wants to thoroughly mix the subjects' names or numbers and 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 the only 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 you 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 & 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 exist on campus. Yet, truly random samples are not always practical. Much psychological research is conducted on college students taking an introductory psychology class, and are required to participate in an experiment for course credit. It is not at all clear whether American introductory psychology college students 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.

### 23.3.2 Correlational studies

A second type of study is the *correlational study* (see Figure 23.2). 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 non-smokers. 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 non-smokers, and randomly assign them to two groups, a "smoking group" and a "non-smoking group." Then the experimenter would simply wait until most of the people in the study have died, and compare the average age and cause of death between the two groups. Because our hypothesis is that smoking causes cancer, it would be clearly 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 non-smokers. 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

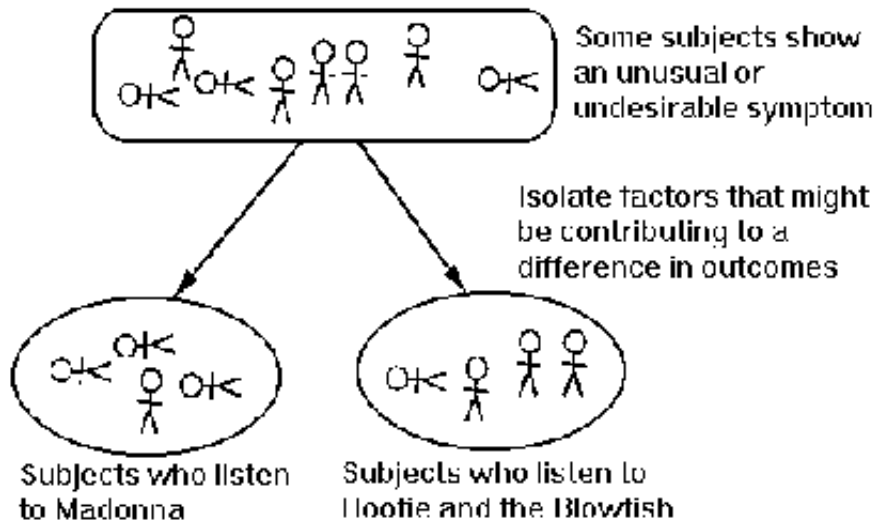
is no causal link between smoking and cancer.

In correlational studies, a great deal of effort is devoted to trying to uncover differences between the two groups studied 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 only prove the presence of something, and 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 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 23.2. 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.

## 2. Correlational experiment



Two possible conclusions:

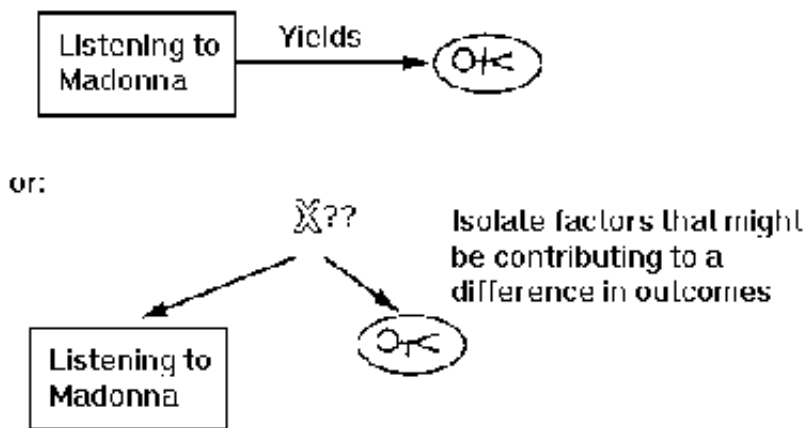


Figure 23.2 In a correlational study, the researcher looks for a relation between two observed behaviors. In this case looking for the relation between untimely death and listening to Madonna recordings.

### 23.3.3 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 the core of other planets, but rather to learn more about the origins of the universe. In psychology, we might want to know the part of the brain that becomes activated when someone performs a mental calculation, or the number of pounds of fresh green peas the average American eats in a year (see Figure 23.3). 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 to 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.

### 3. Descriptive study

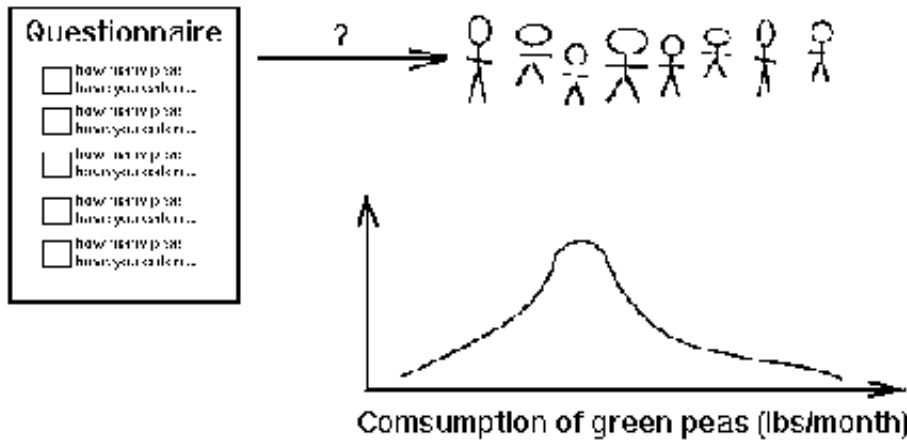


Figure 23.3 In a descriptive study, the researcher seeks to describe some aspect of the state of the world, such as people's consumption of green peas.

## 23.4 Design Flaws in Experimental Design

### 23.4.1 Clever Hans

There are many famous 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 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, 1914/1967). 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 new questions that he had never heard before. In both these cases, Hans continued to perform brilliantly, tapping out the sums or differences to arithmetic problems. In 1904, a scientific commission was formed in order to investigate Hans' 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 seven and three. 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 don't intrude into an experimental situation.

### 23.4.2 Infants' perception of musical structure

In studies of infants' perception of music, infants typically sit in their mother's lap while music is 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 had an innate understanding of musical phraseology.

Are there alternative explanations for the results? Suppose 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 the mothers' perception of the music is masked. This is an example of controlling an extraneous third variable - the mothers' responses to the musical phrases.

### 23.4.3 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 have specified this doesn't make it so! To begin with, monitors have a refresh rate (e.g. 75 Hz) so the "on time" of a stimulus will always be an integer multiple of the refresh cycle ( $1/75 \text{ Hz} = 13.33 \text{ msec}$ ) no matter what you instruct the computer to do. 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 could be delayed by one or two cycles.

It is important, therefore, to always *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 the 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 doesn't mean it is putting out a 1000 Hz signal, you should measure the output signal and verify it for yourself. This is just good science.

## 23.5 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 her findings. The population might be female sophomores at Stanford, or all Stanford students, or all college students in the U.S., or all people in the U.S. 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 2,000 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 of people you are studying. In the extreme, if you are studying a population of people who are so homogenous 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 psychoacoustic experiment, and the phenomenon you are studying is something that you expect to be invariant across people, you only need to use a few subjects, perhaps five. An example of this type of study might 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 error variance, it is a good idea to have at least five to ten subjects in each experimental condition.

### 23.6 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 experimental group would listen to classical music. Of the three control groups: one group would listen to rock music for an equivalent number of minutes per day as the experimental group, although they would not listen to it while they were studying; the second control group would do the same for classical; the third control group 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 ten subjects to each experimental condition, this would require a total of 50 subjects. Table 23.1 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, but there are four music conditions of various types.

Condition		Only while studying	Only while not studying
MUSIC	Classical	subjects 1-10	subjects 11-20
	Rock	subjects 21-30	subjects 31-40
NO MUSIC		subjects 41-	50

Table 23.1 Between-subjects experiment on music and study habits.

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 would 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 non-drug 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. If a subject habituates to a certain sound, it could cause the subject to respond differently to the sound immediately following it than he 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 your 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 ( $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$ ). 720 orders is impractical for a within-subjects design and 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).

### 23.7 Ethical Considerations in Using Human Subjects.

Some particular 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 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 subject has about the procedure or hypothesis. Many psychoacoustic experiments involve difficult tasks, leading some subjects to feel frustrated or embarrassed. A subject should never leave an experiment feeling that they are 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 rather 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.

## 23.8 Analyzing Your Data

### 23.8.1 Quantitative analysis.

#### Measurement error.

Whenever you measure any quantity, there are two components that contribute to the number you end up with: the actual value of the thing you are measuring, plus 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, and 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 for example we are measuring the weight of a sandbag. Formally we would write:

$n \rightarrow \infty, \bar{\epsilon} = 0$ ; [where  $\bar{\epsilon}$  = the mean of epsilon ]

and

$n \rightarrow \infty, \bar{w} = w$  [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 not only encounter measurement error, but 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 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 observed between the two groups can be wholly accounted for by measurement and performance error, or whether a difference of the size you observed 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 only 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, in press). An additional problem with a criterion of 5% is that a researcher who measures twenty different effects is likely to measure one as significant by chance, even if no significant effects actually exist.

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 only differ 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 & Schmidt, 1990; Schmidt, in press). Yet, many people persist in clinging to the belief that the most important

thing to do with experimental data is to test it for statistical significance. There is great pressure from peer-reviewed journals to perform significance testing, 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's 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 (called an ANOVA, a traditional "p-value" test). 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 the *direction* of the difference. The null hypothesis predicts the manipulation would have no effect at all, and that half the subjects should show a difference in one direction and half in the other. The probability of all ten subjects showing an effect in the same direction is  $1/2^{10}$ , or 0.0009 - highly significant. 10 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 significant 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 or the other group will improve your ability to predict their score on the dependent variable, compared to not knowing what group they were in (an example of this analysis is contained in Levitin, 1994a). A good overview of these alternative statistical methods is contained in the paper by Schmidt (in press).

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 variable. Always provide a table of means and standard deviations as part of your report.

### 23.8.2 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.

The first thing one should do with experimental data is to graph it 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 scatterplot. 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. Make sure all your axes are clearly labeled, and don't divide the axis numbers up into units that aren't meaningful (for example, in a histogram with "number of subjects" on the ordinate, the scale shouldn't include 1/2 numbers because subjects only come 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 scatterplot 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, Cleveland, Kleiner & Tukey, 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.

### 23.9 Sources of Experimental Ideas.

There are many ways to generate ideas for an experiment. One is to begin with a theory about how one thing should affect another (such as how music might affect study habits). Another source of ideas is based on observation, such as the observation that people who listen to a particular type of music get sick.

One of the best sources for ideas is to read the reports of previously published studies in scientific journals. By reading about someone else's research, you get a clearer idea of what some of the research problems are. A good article clearly lays out the theoretical issues, provides a background of research on the topic, and reports on studies designed to test certain aspects of the problem. Sometimes the researcher only tackles part of the problem, paving the way for someone else (maybe you) to come along and perform additional studies. Sometimes after reading a report carefully, you might think that the researcher has overlooked an important control, or drawn conclusions that are unwarranted. The search for alternative explanations for an experimental result is one of the more exciting aspects of scientific research -- it is a bit like trying to solve a logic problem or a brain teaser. Was the assignment of subjects truly random? Were the experimental conditions the same, and if not, could the differences have affected the outcome? Reading published studies has another advantage. It helps you to understand and appreciate the types of issues that workers in the field consider important.

Two of the better journals for psychoacoustic research are the Journal of the Acoustic Society of America and Perception & Psychophysics. Other journals publish articles on a wider variety of research topics. The following is a list of recommended journals, and their focus. The first five journals are published by the American Psychological Association.

Psychological Bulletin - "Review" articles on topics of broad interest to psychologists, summarizing and analyzing the body of research to date on a given topic.

Psychological Review - Primarily theoretical papers and in-depth reports of multi-part experiments of general interest.

Journal of Experimental Psychology: General - Experimental reports of general interest.

Journal of Experimental Psychology: Human Perception & Performance - Experimental reports of more specialized research on perception and human performance.

Journal of Experimental Psychology: Learning, Memory, & Cognition - Experimental reports of more specialized research on learning, memory, and other higher cognitive tasks.

Psychonomic Bulletin & Review - Similar to JEP: General or Psychological Review, but published by the Psychonomic Society, featuring experimental and theoretical papers.

Music Perception and Psychology of Music - experimental reports in music perception and cognition.

Psychological Science - published by the American Psychological Society, featuring reports of interesting experiments on a variety of topics, similar to JEP: General or Psychonomic Bulletin & Review.

Current Directions in Psychological Science - also published by the American Psychological Society, featuring brief reports of interesting experiments, usually without in-depth coverage of methods and statistical analyses.

Science, Scientific American and Nature - articles of general interest or importance covering all topics in the natural sciences, life sciences, behavioral sciences, and some engineering. Their articles on psychoacoustic studies are generally excellent, but few such articles are reported in any single year, and they do not typically report in detail on experimental methods or analyses.

Once you have an idea for a topic, it is important to perform a literature search. By reading previous reports on the topic in which you are interested, you can see what other experiments have been done on your topic. It is possible someone had the same idea as you, and already ran the experiment. What is more likely, though, is that someone had a similar idea, but not identical to yours, and you can learn a great deal by reading about how they approached the problem. This can also be a good source of guidance for experimental design.

A good place to start a literature search is "PsycINFO," an on-line database available through many university libraries. It indexes over 1,200 journals in psychology and associated fields, with coverage from 1984 through the present. Its sister database, "PsychLit" is available on CD-ROM at many university libraries, and offers expanded coverage, going back to 1978 and including book chapters as well as journal articles. Once you find one article or book chapter related to your topic, you can use its bibliography to direct you to other papers.

### **23.10 Special Considerations in Psychoacoustic Research**

There are particular problems that arise in psychoacoustic research unique to the field and should be considered when designing an experiment.

1) Perceived loudness is frequency-dependent. That is, given two auditory signals at different frequencies, and with equal power (or amplitude), there will be systematic and predictable differences in their perceived loudness as a function of frequency. The Fletcher-Munson curves (Fletcher & Munson, 1933), available in most acoustic textbooks, illustrate the way in which loudness varies with frequency. If your experiment

involves presenting tones of different frequencies to your subjects, you will need to remove perceived loudness as a variable. The most rigorous way to do this would be to compute equal loudness curves for each subject and adjust the power of your signal accordingly. A more common (and practical) solution is simply to vary the loudness of different signals randomly. This effectively removes loudness as a source of variance in your data.

2) Reverberation. In studies of perceived durations or locations of different tones, the room in which you are playing the signals can add a significant amount of delay to the signal, affecting subjects' judgments. One solution is to conduct the experiments in an anechoic chamber. A more practical solution is to use headphones to present the stimuli. Be certain to use "over-the-ear" headphones that encompass the entire ear, so that the sound is less likely to escape into the room and reverberate.

3) Hysteresis. Hysteresis refers to the failure of a system to return immediately to its original state following stimulation. Hysteresis effects are not confined to psychoacoustic work but appear in studies involving all psychophysical domains. In acoustic studies, there are two possible sources of hysteresis effects: the physical equipment being used to create sound, and the human auditory system. Analog physical equipment, such as oscillators, do not always produce the same output with identical settings of the dial, resulting from hysteresis and "slop" in the physical components. Similarly, the human auditory system does not always return to its resting state following the presentation of a given stimulus. Suppose that the subject's task is to tune a variable oscillator to match a tone they have held in memory. If the subject adjusts the oscillator upward from a low tone, they might settle in on a different frequency than if they adjust the oscillator downward from a high tone. Furthermore, there may be a range of frequencies that sound the same to the subject. We might talk about this as "slop" in the setting of the oscillator. It is important, therefore, in experiments in which the subject makes adjustments of some sort, that you average across several trials, and that you measure with adjustments from both above and below the target point.

4) Echoic memory. In studies of auditory memory and perception, it is often necessary to ensure that a given stimulus tone is masked, or erased, from the subjects' short term memory, in order that the judgment of the subsequent tone is not influenced by memory of the first. The most effective way to erase memory of a pitch is to play a sequence of random tones. More specific details about auditory masking are contained in Butler and Ward (1988).

### 23.11 Checklist: The ten stages of a research project

Research is a collaborative process. Do not be afraid to discuss your ideas with other people, even if you are not clear about them yourself. A research project is not like a closed-book exam in which you are not allowed to get help. On the contrary, it is an opportunity for you to interact with people who can be collaborators in the process of reasoning through a scientific problem, and who share the same intellectual interests and curiosities as you do. Accordingly, the following checklist emphasizes the importance of talking to people - fellow students and instructors - throughout the process of conducting a study.

**1. Get an idea.** You may already have this, or you can consult previously published reports, because many papers include a section called "directions for future work."

**2. Perform a literature search.**

**3. Talk to people.** By this stage of the project, you have an idea of what you want to study, and these

colleagues can help you to figure out if (a) the idea is theoretically interesting, (b) it has been done (you might have missed it in your literature search), and (c) the experiment is actually "do-able," that is, if you can study the problem using the time, methods, and materials resources that are available to you.

**4. Set up a timeline.** Most students undertaking their first independent project dramatically underestimate the amount of time it will take. Setting up computer programs or aparati to present stimuli and to analyze data can involve some setbacks. The actual experiments might not run smoothly. The subjects might not show up for the scheduled session, and so on. It is important to consult with your research advisor to come up with a reasonable estimate for how long the study will take, and to create a time-line of completion dates for each phase of the project.

**5. Design the study.** Often you can use the procedures in a previous study to help give you ideas for equipment, conditions, and stimuli to use. During this phase of the project, specify how many subjects you will use, what their task(s) will be, and how you will measure and record their responses.

Once you have designed the study, and before you have tested any subjects, it is a good idea to begin writing your paper. At this point in the project the literature review is fresh in your mind, so you can easily write the introduction and methods sections, as well as the references section. A brief account of the hypotheses being tested and the procedures used will be needed for human subjects approval. A description of the proper format of a research paper is beyond the scope of this chapter, but examples can be easily obtained in the journals listed earlier, or in the Publication Manual of the American Psychological Association (American Psychological Association, 1994).

**6. Run a pilot study .** Now that the study is designed, run a "pilot study" with a few subjects just to see how everything works. This gives you a chance to test the equipment, practice what you will say to the subjects, and make sure that the task makes sense to the subjects. It also lets you see how long the experiment will take. Analyze the pilot data to see if they make sense. In the short ten weeks we have in the Psychoacoustics course at Stanford, a pilot study is often as far as our students can get, but it forms the foundation for further work.

**7. Run the study.** Be sure to keep accurate records for each subject: the date and time the subjects was tested, the subjects' name, age, sex, and so on. Be sure to ask the subjects about any variables that might affect their performance. In a hearing study, do any of your subjects have a cold? In a vision study, do any of your subjects wear glasses?

**8. Analyze the data.**

**9. Talk to people.** After you have performed your analyses, it is a good idea to talk to fellow students and advisors again. Do your findings suggest follow-up studies? Do they suggest an explanation due to experimental artifacts? Have you overlooked an important analysis?

**10. Write up the results.**

A good report will include an introduction to the problem, a discussion of previous studies on the topic (you should have done all this in step 4 above) and a lucid presentation of your findings. In both the results and discussion sections, discuss the data qualitatively. Discuss what you think the data mean. Also discuss possible alternative explanations for your findings, and tests that might be done in the future to control for artifacts. Be careful in your report not to draw conclusions that are not supported by the data, and not to

overgeneralize your findings.

### 23.12 Coda: An Example of a Class Study

Anne Stern was an undergraduate student taking the Psychoacoustics course at Stanford CCRMA in 1993. In addition to being an interesting and clever experiment, her final project illustrates clearly some of the design issues we have reviewed. Anne's question was this: if one were to ask people to sing a musical note with no further instructions, what note would they sing? Across a large number of people, will the productions tend to cluster around middle "C"? Around "A"? Or will the notes be uniformly distributed? Her idea was to simply catch people in White Plaza (a central point on the Stanford Campus, between The Bookstore and the Coffee House) and ask them to sing the first tone that popped into their heads, using the syllable "la." This is an interesting question for memory researchers who wonder whether we have an internal template for musical tones, and it is an interesting question for music psychologists. This concluding portion of this chapter is a class discussion of her project presented in dialog form.

[Daniel Levitin]: What are some of the things she needs to think about in designing the experiment?

[Student]: Different vocal ranges?

[Daniel Levitin]: Right - people are going to have different vocal ranges. We can ignore octaves, as they do in absolute pitch research. Because, really, what we are interested in is pitch class, or chroma. What else?

[Student]: They might have just heard some music and that could influence the tone they sing.

[Daniel Levitin]: Yes - White Plaza often has live bands or a boom box playing, and this could skew the results; if subjects' productions cluster around a tone, it could be because they were all influenced by some external source.

[Student]: The method of recording and playback might be variable...the tape recorder speed could change as the batteries wear down.

[Daniel Levitin]: Right - this is a very important point. The ideal way to record would be digitally; this way, we can be sure that the pitch stays constant even if the motor fluctuates. What else might we want to control for?

[Student]: Students in White Plaza might not be representative of students in general...there might be a particular type of student who goes to White Plaza, and they may differ systematically from other students.

[Perry Cook]: Also the time of day is a confound - people who sing at ten in the morning might sing a different pitch than they would sing at four in the afternoon when their vocal chords are warmed up.

[Daniel Levitin]: Both of these are important. Ideally, you would like to pick student numbers at random from a bin, so that you knew nothing about the person ahead of time. There is always the problem in psychological research that you have not selected people in such a way that you can generalize your findings. Most of the research done in the Psychology Department at Stanford, whether it's on cooperative attitudes, memory, social behaviors, or decision making, is done on Introductory Psychology students. Now it is possible that Introductory Psychology students are representative of students-in-general on campus, but it is also possible that they are not. In fact, it turns out that students enrolled in Introductory Psychology at

Stanford tend to be more depressed than students in general. This may or may not have an affect on your experiment.

Well, Anne collected her data and here is what she found (see Figure 23.4).

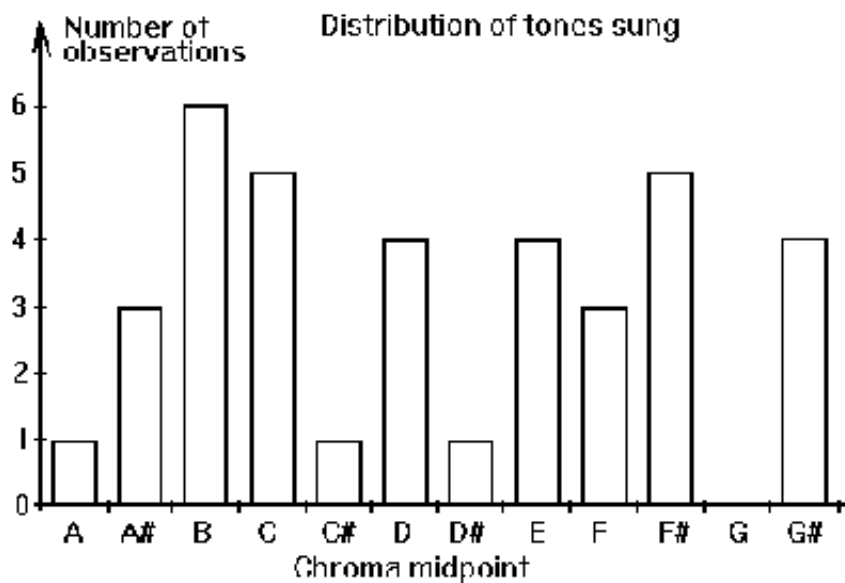


Figure 23.4 Data on tone production from Anne Stern's class project.

[Daniel Levitin]: Notice that the modal response was to sing "B" and nobody sang "G". The research question is whether there is a preferred tone, or whether these results are uniformly distributed. What do you think just by looking at it?

[John Pierce]: Uniform.

[Daniel Levitin]: That's right. Even though nobody sang "G" and more people sang "B" than any other tone, these minor variations from uniformity are what we might expect by chance with such a small sample. By performing the Rayleigh test - a goodness of fit test - we can conclude that the distribution we observed is more likely to have arisen from a uniform distribution than a unimodal distribution with "B" as the mode. (For details on the Rayleigh test, see Levitin, 1994b). In other words, suppose that we hypothesize that the true state of the world is that the distribution of tones is uniform. What the statistics are telling us is that this distribution, even with its non-uniformities, is likely to have arisen as a result of sampling error, not as a result of the state of the world being different than we think it is.

[Student]: What would happen if you had a larger number of subjects?

[Daniel Levitin]: If our statistical inferencing is correct, if we had ten times as many subjects, the frequencies of tones produced in each category should start to even out.

[John Pierce]: In a rather different experiment, Diana Deutsch found a difference between English people and Americans in their perception of ambiguous Shepard tones (Deutsch, 1991; Deutsch, 1992). And so one might, in an experiment like this, ask what country the subjects came from.

[Daniel Levitin]: Yes - that would have been an important control in this study. Anne might have restricted

her subjects to Americans, or she might have asked them about their country of origin.

I like this study as a course project because it addresses an interesting question, it allowed Anne to gain experience in a number of different techniques, and it was possible to do a good job on the study in the few weeks that she had.

[John Pierce]: It's important to note the difference between *demonstrations* and *experiments*. A demonstration, such as of the *precedence effect* or of *auditory streaming*, is something you prepare and play for a large number of people, most of whom perceptually experience the effect. For the purposes of this class requirement, demonstrations are an acceptable project. If you find some way of demonstrating an effect, either one that was unknown or one that was previously known, and you can demonstrate it in a way that is very clear and overwhelming, that represents a significant contribution.

[Daniel Levitin]: Diana Deutsch's tri-tone paradox is an example of a demonstration.

[Brent Gillespie]: Another important point in experimental design is to ask a research question in terms of a hypothesis that is falsifiable. You want to be able to run an experiment that can disconfirm a very specific hypothesis.

[Daniel Levitin]: Yes. The best situation is if you can design an experiment such that either way the data come out, you have an interesting finding. Anne's study is an example of that - whether she found a uniform distribution or a unimodal distribution, either result is interesting. A less interesting study is one in which you only have something interesting to report if the data come out one particular way.

[Perry Cook]: How do you decide what kind of response you want from your subjects, forced-choice or free-response?

[Daniel Levitin]: As the names imply, in a *forced-choice* experiment, the subject knows all the possible responses and they just choose one; in *free response* the subject answers without any specific guidance from the experiment or the experimenter. In a *memory study*, or in a *psychophysical identification* study, you often want to use free-response in order to get an unbiased answer. But forced-choice responses are easier to code (or evaluate) because the responses are constrained. There's no easy answer to this question and psychoacousticians fight about this all the time.

[John Pierce]: This is tough stuff. I want to convey to all of you that you should consult with us on your project so that we can steer you away from undertaking too much. Students usually tackle too big a problem to do in a term project. What *is* practical in Psychology 151/Music 252? Demonstrations show that some things are so SO that people can be convinced in a very informal way. You said earlier that it is not difficult to think up good psychology experiments, but I think it is very difficult to think up ones that are worth doing and can be handled within your demands for subject time, variety, and methodological cautions. This is why I like demonstrations. And Anne's study is good.

[Daniel Levitin]: It is important for you to think about experiments early in the term so that you can change your mind if need be and still have time to do something worthwhile.

## Author's Note

This chapter benefitted greatly from comments by Perry Cook, Lynn Gerow, Lewis R. Goldberg, John M. Kelley, and John R. Pierce.

As of 8/1/96, address correspondence to: Daniel Levitin, Behavioral Science Laboratory, Interval Research Corporation, 1801C Page Mill Road, Palo Alto, CA 9430. Phone: (650) 842-6236. E-mail: levitin@interval.com

## 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 ed.). Washington, DC: American Psychological Association.
- Butler, D., & Ward, W. D. (1988). Effacing the memory of musical pitch. Music Perception, *5*(3), 251-260.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). Graphical methods for data analysis. New York: Chapman & Hall.
- Cleveland, W. S. (1994). The Elements of Graphing Data. (Revised ed.). Summit, NJ: 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 ed.). Mountain View, CA: Mayfield Publishing Company.
- Daniel, W. W. (1990). Applied nonparametric statistics. (2 ed.). Boston: PWS-Kent.
- Deutsch, D. (1991). The tritone paradox: An influence of language on music perception. Music Perception, *8*, 335-347.
- Deutsch, D. (1992). The tritone paradox: Implications for the representation and communication of pitch structure. In M. R. Jones & 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., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. Journal of the Acoustical Society of America, *72*, 82-108.
- Hayes, W. (1988). Statistics. (Fourth ed.). New York: Holt, Rinehart and Winston.
- Hempel, C. G. (1966). Philosophy of natural science. Englewood Cliffs, NJ: Prentice-Hall.
- Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.
- Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences. (2nd ed.). Pacific Grove,

CA:: 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.

Levitin, D. J. (1994b). Problems in applying the Kolmogorov-Smirnov Test: The need for circular statistics in psychology (Tech. Report #94-07): University of Oregon, Institute of Cognitive & Decision Sciences.

Schmidt, F. L. (in press). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods.

Shaughnessy, J. J., & Zechmeister, E. B. (1994). Research methods in psychology. (Third ed.). New York: McGraw-Hill.

Stern, A. W. (1993). Natural pitch and the A440 scale (Unpublished report): CCRMA, Stanford University.

Watson, J. B. (1914/1967). Behavior: An introduction to comparative psychology. New York: Holt, Rinehart and Winston.

Zar, J. H. (1984). Biostatistical analysis. (Second ed.). Englewood Cliffs, NJ: Prentice-Hall

[Go Back to Daniel Levitin's Home Page.](#)