Adv Physiol Educ 31: 82–92, 2007;
doi:10.1152/advan.00033.2006.

Teaching simple experimental design to undergraduates: do your students understand the basics?

Sara M. Hiebert

Department of Biology, Swarthmore College, Swarthmore, Pennsylvania

Submitted 13 May 2006; accepted in final form 4 September 2006

Hiebert SM. Teaching simple experimental design to undergraduates: do your students understand the basics? *Adv Physiol Educ* 31: 82–92, 2007; doi:10.1152/advan.00033.2006.—This article provides instructors with guidelines for teaching simple experimental design for the comparison of two treatment groups. Two designs with specific examples are discussed along with common misconceptions that undergraduate students typically bring to the experiment design process. Features of experiment design that maximize power and minimize the effects of interindividual variation, thus allowing reduction of sample sizes, are described. Classroom implementation that emphasizes student-centered learning is suggested, and thought questions, designed to help students discover and name the basic principles of simple experiment design for themselves, are included with an answer key.

controls; randomization; Student's *t*-test; teaching

DO YOUR STUDENTS understand the basics of experimental design? Do they really know what a control is? Do they know which statistical test to use when comparing two treatment groups, and do they know which groups to compare? In the course of teaching an intermediate level course in animal physiology over the past decade, I have become aware that these concepts are poorly understood by most undergraduates, and that even an entire 3-h class session devoted to this subject is rarely enough to completely replace the deep-seated misconceptions that many students hold about even the simplest experimental designs. Judging from my interactions with high school science teachers, these misconceptions are not limited to undergraduate students; they subsequently persist both in graduate school and in published physiological research, weakening the rigor and potentially increasing the cost of investigation (4, 5,16, 28). In this article, I describe the content of a 3-h laboratory class in which students are asked to design an experiment to determine whether chicken embryos are endothermic or ectothermic. The specific procedures and theoretical basis of this experiment are described in greater detail in the companion article "Are chicken embryos endotherms or ectotherms?" (10). However, the general concepts, discussion plan, and thought questions are applicable to any experiment in which the effects of two treatment conditions are being compared.

Most students know that an experiment should contain a control, but many find it difficult to define exactly what a control is. A high school science teacher in a recent laboratory workshop said that, in general, a control is "the animal doing nothing," by which she meant that the animal is in its home cage where it is not exposed to any features of the experimental treatment, not even the features for which the experiment needs to be controlled. Students often have difficulty both in verbalizing the general way in which a control is related to experimental treatments as well as in specifying the appropriate control(s) for a particular experiment. Many also believe that there must always be a specific treatment group that is designated as a control and that the only permissible way to create these different treatment groups is by random assignment. Students who have taken a statistics course may know more about statistical tests, but they don't always remember how to apply them to an appropriate experimental design. For example, once students are introduced to the statistical power of the paired *t*-test, they are eager to design experiments that use this test but frequently fail to incorporate adequate controls. A typical experiment design proposed by students is one in which levels of some physiological variable are measured first to get a baseline reading; these baseline measurements are then compared with measurements on the same individuals in response to a later treatment. Students are surprised to learn that such an experiment cannot, in fact, tell us whether the experimental treatment had any effect.

As those of us who are not full-time statisticians know from personal experience, it is far more effective to learn experimental design and statistical analysis in the context of an actual experiment. I find that the most effective incentive is an experiment that students must design to address a question to which they do not know the answer. Thus, I use a student-designed experiment on chicken embryo metabolism as an opportunity to teach the basics of simple experimental design and correct some of these misconceptions.

Classroom Implementation

I use the design process described here as the second 3-h laboratory session in a three-session module in which students learn laboratory techniques (*session 1*), design an experiment (*session 2*), and perform the experiment (*session 3*). Many implementations are possible, however, and some alternatives are discussed in Ref. 10.

The overall goal of the 3-h experiment design session is to reduce teacher-centered instruction and engage students in active learning as much as possible (3, 20, 21). I begin the laboratory session by providing a brief overview of the three types of *t*-tests (two-sample, paired, and one-sample *t*-tests). Most students entering the course in which I use this laboratory exercise have had previous practice with at least two of these *t*-tests. Armed with this information, students in each laboratory section (12 students) are asked to break into groups of two or three students. Each group is asked to arrive at a single design that meets the following criteria:

1. The experiment must use measurements of respiration rates ($\dot{V}O_2$) at different temperatures to determine whether 17-day-old chicken embryos are endotherms or ectotherms.

2. The experiment must be properly controlled.

3. The results should be analyzed with one of the *t*-tests we have discussed.

4. The experiment must be completed within one 3-h laboratory session with the available equipment.

During the student-centered design process, I am prepared to provide handouts with the relevant additional information that the students will need to design a realistic biological experiment. In the case of the chicken embryo metabolism experiment, such information includes the safe range of temperatures for avian embryos (abstract from Ref. 26; 16–41°C) and the time that might be required between measurements for temperature equilibration (egg cooling curve, Fig. 2 in Ref. 10). The rationale for giving this information to students when they ask for it rather than at the beginning of the design session is that in real experiments, investigators are not presented with all the information they need at the outset. Students who learn to think logically about what they need to know before proceeding, and who can formulate specific questions that they can answer by consulting the instructor or literature, are better prepared to launch their own independent projects later in the course.

After each student group has finished this task, one group is chosen to share its design with the class. The design process progresses as this first design is modified to (or replaced by) a mutually agreed upon design in a discussion moderated by the instructor. As a starting point for the group discussion, it is actually instructive to choose a design with flaws, so that the class can engage in a discussion about how to improve the design to meet the requirements of the experiment. In the process of revising the design, I ask students to come up with some "rules" (general principles) for experimental design that they can apply to future experiments as well.

The thought questions presented at the end of this article can be used in a variety of ways to support the design process. They may be assigned, for example, as preparatory or followup homework. However, in my laboratory, I prefer students to work on at least some of these questions in small groups during the class period, when I am available for discussion. This means that relevant student questions can be addressed immediately and can serve as a learning experience for everyone in the class, not just the student who posed the question. Each time I lead an experimental design session, the discussion takes a slightly different course. I assign individual questions when the concept covered by that question becomes relevant to the discussion; students break into small groups to answer the question(s) before reviewing the answers as a class and moving on to the next stage of the design process.

In the weeks following an intensive session on experiment design, do not be surprised if your students need further reinforcement to establish firmly the concepts introduced in this lesson. Practice is essential. In my course, students write a strong-inference protocol after each design session. Writing this short but important document helps students to process what they have discussed in class, prepares them for performing the experiment, and provides ideas and text for the final laboratory report. For a general discussion of the strong-inference protocol and specific examples for the chicken embryo metabolism experiment, see the companion article (10).

Classroom outcomes. To ensure that students have engaged with the concepts on their own before participating in the small-group discussions, they are asked to bring a draft experimental design to class on the day that we design our chick embryo metabolism experiment. Although I do not grade these draft designs, I collect them because they serve as a useful gauge of students' prior knowledge and proficiency in experimental design. In a typical laboratory section of 12 students, only 1 or 2 students will have produced a well-designed experiment in which they have explained controls explicitly. Of the remaining students, roughly one-half will have designed an experiment with unclear controls, with wording such as "of course we would have to include controls" but without explaining what treatment the controls would undergo. The remaining designs typically incorporate obvious flaws, such as measuring all embryos at one temperature and then all of the embryos at another temperature.

Despite the need for periodic reinforcement of the concepts they have learned, students who have completed the 3-h experimental design session for the chicken embryo metabolism laboratory are better able to design future experiments and to do so faster. Three weeks after the chicken embryo metabolism exercise, the entire laboratory section goes on to design a second experiment concerning the effect of temperature on lizard locomotion. This time, it takes only ~20 min of class time for students to design a well-controlled experiment that all members of the class can readily agree on. These two laboratory exercises are designed to prepare students for an independent project that they will undertake individually or in small groups in the second half of the semester. As with the lizard locomotion exercise, students appear readily able to apply the general concepts that they have learned from the chicken embryo experiment to other experimental problems. An example is provided by a student who approached me at the very beginning of the semester with the concern that she had never designed an experiment before and that she was therefore worried about her ability to complete the independent project that would be required of her later in the semester. When it came time for the independent project, however, she formulated an interesting question and designed a well-controlled experiment without any instructor assistance.

Experimental Designs for Comparing Two Treatment Groups

Two types of experimental designs typically emerge from the classroom discussion during a design session.

Design 1: control group. The general design of this experiment (Fig. 1*A*) calls for dividing subjects into two groups, one of which is designated the control. Baseline measurements on all animals at a starting condition are used to create balanced treatment groups (see *Creating treatment groups*). The experimental group is then exposed to the experimental treatment, while controls are maintained at the starting condition under which baseline measurements were made. A second set of measurements is then made on all animals. The data are analyzed by comparing this second set of measurements between control and experimental animals with a two-sample *t*-test.

EXAMPLE FOR THE CHICKEN EMBRYO METABOLISM EXPERIMENT. The rate of oxygen consumption ($\dot{V}O_2$) of all 12 eggs is measured at one ambient temperature (usually 38°C, the normal incubation temperature for chicken eggs). Eggs are then divided into balanced groups (see *Creating treatment groups*) based on baseline $\dot{V}O_2$. A randomly chosen one of these groups

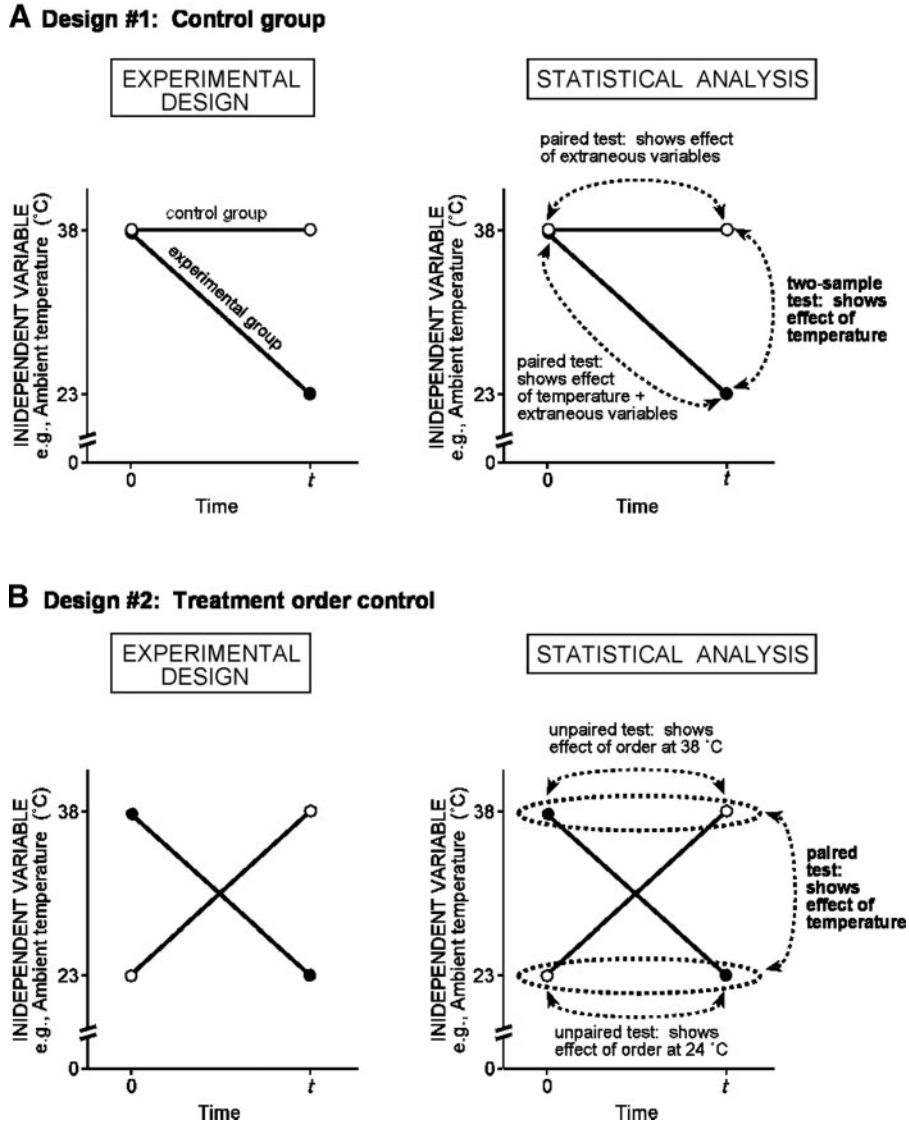

Fig. 1. Graphical depiction of two experimental designs for comparing two groups in a test for the effect of an independent variable at two levels (X_1 and X_2) when measurements of the dependent variable is made at two different times (*time 0* and *time t*). Note that students often confuse figures that show experimental treatments (such as those here) with figures that show the results of the experiment; the latter would be depicted as the dependent variable (*Y*) graphed against the two levels of the independent variable (X_1 and X_2), as in Figs. 3, 4, or 6*A*. *Design 1* (*A*) uses a control group to control for time-dependent extraneous variables such as previous handling, habituation, learning, and developmental age, whereas *design 2* (*B*), also known as a crossover design, uses treatment order to control for time-dependent extraneous variables without designating any one treatment group as the control group. Each design is shown with the statistical analyses that are possible and that are appropriate (shown in bold) for determining the effect of the independent variable of interest. In the specific context of the chicken embryo metabolism experiment, the independent variable is ambient temperature, with suggested values of $X_1 = 38°C$, $X_2 = 23°C$, and *time t* = 90 min for the second measurement of *Y* [rate of oxygen consumption ($\dot{V}O_2$)]. In *design 1*, *comparison A* (paired *t*-test) tells us the effect of the entire treatment on the dependent variable ($\dot{V}O_2$). *Comparison B* (two-sample *t*-test) tells us the effect of the independent variable (temperature) alone; it is the only comparison that needs to be tested statistically to answer the question. *Comparison C* (paired *t*-test) tells us about the effects of the extraneous variables for which we are controlling. In *design 2*, the question is answered by using a paired *t*-test to compare the values of the dependent variable ($\dot{V}O_2$) of all embryos between the two levels of the independent variables (38 and 23°C). The dashed-line circles indicate that the test compares all $\dot{V}O_2$ at 38°C with all $\dot{V}O_2$ at 23°C regardless of the order in which the data were obtained. Comparing the $\dot{V}O_2$ between the two treatment groups at either temperature (with a two-sample *t*-test) tells us whether there is an effect of treatment order on $\dot{V}O_2$ at that temperature.

(designated the experimental group) is then placed at a lower temperature (room temperature, ~23°C, is convenient). Control eggs are handled similarly but returned to 38°C. After 90 min of equilibration at these two temperatures, the $\dot{V}O_2$ of all experimental and control eggs is measured again, and a two-sample *t*-test is used to compare these measurements between the experimental eggs at 23°C and control eggs at 38°C.

Design 2: treatment order control The general design of this experiment (Fig. 1*B*), often referred to as a crossover design, calls for each subject to be measured under both conditions, with one-half of the subjects experiencing the conditions in reverse order. A paired analysis is used to compare the measurements obtained in the two conditions. Because the analysis considers only the differences between the two measurements from each individual, this powerful design eliminates the "noise" resulting from differences between individuals.

EXAMPLE FOR THE CHICKEN EMBRYO METABOLISM EXPERIMENT. The $\dot{V}O_2$ of each embryo is measured at each of two ambient temperatures (38 and 23°C are convenient, as above). Six of the embryos are measured first at 38°C and then at 23°C, while the other six embryos are measured at the same times but in the reverse order of temperature treatments. As in *design 1*, the two measurements are separated by 90 min to allow for thermal equilibration.

General Considerations for Both Designs

Regardless of which design is ultimately chosen, several important elements should be included in the classroom discussion. Note that a discussion of statistical analysis cannot wait until after the data are collected. Rather, the experiment is designed specifically with statistical analysis in mind, to insure that the data will be suitable for analysis and that the data will actually answer the question addressed by the experiment. See the companion article, "The strong-inference protocol: not just for grant proposals" (9), for further discussion and specific examples.

Controls. In both designs, we measure $\dot{V}O_2$ of some or all of the eggs first at one temperature and then at another. It is important to remember, however, that because the two measurements did not occur at the same point in time, temperature is not the only variable that the embryos will have experienced

differently in the two measurements. Other variables that differ between measurements taken at different times include the *1*) total amount of handling, *2*) time since first handling, *3*) learning, *4*) time of day, and *5*) when the time difference between measurements is large and/or the animals are developing rapidly, stage of development will be different for measurements taken at different times of the day.

Thus, we need to *control for* the effects of the time-dependent variables in which we are not interested (the extraneous variables) to gain information about the independent variable of interest (temperature). There are two methods for doing so. The first is to run a parallel control group that receives exactly the same treatment as the experimental group except for the variable of interest (*design 1*; Fig. 1*A*). In this case, all of the subjects experience exactly the same amount of handling, are being measured again the same number of hours after first handling, and are being measured at the same times of day and at the same developmental stage. Therefore, the only variable that differs between the control and experimental groups at the second measurement is temperature, and if we compare the $\dot{V}O_2$ between the two groups at the second time point, any differences should be due to temperature alone. We thus define a control group as a group of animals that receives treatment identical to that of the experimental group except for the independent variable of interest.

The second general method for controlling for extraneous variables is to measure each individual in both experimental conditions (here, high and low ambient temperature) but to randomize the order in which subjects experience these conditions so that any directional effects of the four extraneous variables are averaged and therefore cancel one another out when $\dot{V}O_2$ is compared between the two temperatures (*design 2*; Fig. 1*B*). In this design, there is no group designated as a control; rather, the overall design of the experiment controls for the extraneous variables.

Creating treatment groups: is random assignment always best? Ask any student how individuals should be separated into treatment groups, and he or she will dutifully reply, "Randomly." Yet, most students are intuitively aware that when a group of organisms is divided into two treatment groups, it would not be a good idea to put all the females in one group and all the males in the other or all the largest individuals in one group and all the smallest in the other. By assigning subjects to treatment groups in a way that avoids creating groups with obvious differences, however, they would be violating the edict of random assignment. Which procedure, then, is the correct one and how can we decide which to use?

First, it must be understood that random assignment is not the goal. The goal is to avoid creating unbalanced samples. Under many conditions, random assignment is the preferred method for achieving this goal (e.g., Refs. 4 and 16). When we have little or no information about the individuals we plan to test, random assignment is the only way to proceed. However, because random assignment is random, occasionally this method will result in unbalanced samples (e.g., one group might contain significantly more males or significantly larger individuals than the other). Even if unintentional, such differences pose problems for interpreting the results of the experiment later on. Consider an experiment in which birds are randomly assigned to two treatment groups. One group is given a hormone treatment, and the metabolic rates of all the animals are then measured to determine whether the hormone has any effect on metabolic rate. If random assignment had, by chance, resulted in two groups that differed substantially in average body mass or sex ratio, this experiment would be unable to separate the effects of mass, sex, and hormone treatment, all of which are known to affect metabolic rate. In other words, the effect of the hormone is *confounded* with the effects of mass and sex in this experiment. Treatment groups can validly be rebalanced after data have been collected (22), but this procedure is undesirable because it reduces sample size, and thus statistical power, and necessitates a more complex subsequent statistical analysis that is beyond the scope of students just beginning to learn the basics. It is far better to balance groups at the beginning of the experiment, and this is a practice we should be teaching our students explicitly so that they too can get the most from every one of their experiments.

Whether or not an investigator knows about or routinely practices preexperiment balancing seems to be related largely to the particular organisms the investigator studies. When inbred strains of rat of a particular sex and age are being studied, for example, the animals are so similar genetically and physiologically that random assignment to treatment groups is highly unlikely to produce groups with systematic differences. Experimenters who work with wild animals, on the other hand, tend to balance treatment groups as a matter of course, particularly when sample sizes are limited because the animals are rare, difficult to catch, or require elaborate holding facilities. In general, randomized group assignment is preferred when sample sizes are large or when the experimental organisms are very similar. When sample sizes are small and there is considerable interindividual variation, however, creating balanced groups greatly reduces the chance that a variable other than the intended independent variable is responsible for the effects observed. Many students are in fact already familiar with the general principle that the risk of selecting a sample that does not represent the population increases as sample size decreases. They may know this phenomenon as "sampling error," and may have encountered it in the form of the founder effect in the evolution of small, isolated populations. This is an excellent opportunity to point out that the same principle is at work here.

Far from being an "advanced" technique unnecessary for undergraduates to learn, balancing groups at the outset of an experiment is a basic procedure that can improve the odds of success for any experiment involving small numbers and/or animals with large interindividual differences. Just as it is important to inculcate students with the principles of ethical animal treatment at the earliest possible stage, it is important to train them to design the most effective experiments with the smallest feasible sample sizes. Treatment group balancing is one of the important tools that we can use to comply with the 3 Rs of research–reduce, refine, and replace–as mandated by the Animal Welfare Act and expanded on in the National Institutes of Health *Guide for the Care and Use of Laboratory Animals* (13).

WHICH VARIABLES SHOULD BE USED TO BALANCE TREATMENT GROUPS? If we are interested in the effects of an independent variable on $\dot{V}O_2$ but we know that individuals vary considerably in their baseline metabolic rates, then baseline $\dot{V}O_2$ would be the most useful measure for dividing experimental subjects into balanced groups. However, in many experiments, baseline measurements are never made; instead, it is assumed that

random assignment will create evenly matched treatment groups. If this assumption is met, such an experimental design is perfectly adequate. The results of such experiments appear simply as the data that are collected in the measurement at *time t* in experimental *design 1* (Fig. 1*A*).

There are some experiments for which baseline measurements of the variable being studied are not practical. For example, it might be useful to know what an animal's baseline response to an injection of endotoxin (lipopolysaccharide) is before assigning animals to treatment groups in a study of how the response to lipopolysaccharide changes with age or is affected by a particular pharmaceutical agent. However, because animals habituate rapidly and strongly to lipopolysaccharide treatment, the baseline test would severely blunt or possibly even obliterate the response of interest in the experimental test. Even in such cases when information about the variable of interest is not available, it is to the experimenter's advantage to increase the probability of creating balanced groups by considering other independent variables that are likely to have an impact on an animal's response in an experiment. Body mass (1, 2, 5, 19), sex (5, 6, 14, 15, 27), age (e.g., 5, 12, 25), and strain or population (4, 24, 25) are four of the most common, most easily determined, and potentially most important variables.

HOW DO WE KNOW WHEN TREATMENT GROUPS ARE BALANCED? Mean and variance in any trait potentially affect the mean and variance of the independent variable, which in turn contribute to the outcome of statistical comparisons. Thus, both parameters should be balanced for each variable being considered. A conservative rule of thumb is to create groups with similar SEs (or variances or SDs) in which the ranges for means ± SE in the treatment groups overlap one another (8). For instructions on how to set up and use a spreadsheet to create balanced groups, see Fig. 2. Once balanced groups have been created, the groups should be randomly assigned to the experimental and control treatments (7). Note that the initial measurements used as the basis for balancing groups (e.g., $\dot{V}O_2$) should be collected in as unbiased a way as possible; alternatively, potentially biased variables can be added to the list of variables to be balanced. Consider, for example, an experiment in which eggs arriving in one carton are numbered *1–12* and those arriving in another carton are numbered *13–24*; the $\dot{V}O_2$ of the eggs in the first carton is measured first and $\dot{V}O_2$ of the eggs in the second carton is measured second. In this case, carton number should be one of the variables balanced.

Statistical analysis with Student's t-test. When designing their experiment, students are instructed that they may use no more than one *t*-test to analyze the results. Each time a *t*-test is performed and a maximum *P* value of 0.05 is used as a criterion to determine whether the difference between the two samples is statistically significant, we are accepting a 5% chance that the difference we observe is in fact produced by chance and does not reflect a real difference between the two populations we are comparing. Multiple comparisons requires the use of ANOVA to control for the additional uncertainty introduced by accepting a 5% chance of error (or other limit) for each comparison. The *t*-test may be thought of as a special case of the ANOVA for comparisons between just two groups.

DESIGN 1. The only *t*-test required is a two-sample comparison between the experimental and control groups at the second measurement, as discussed above. In fact, *no other comparison* will isolate the effect of the independent variable on the dependent variable (see Fig. 1*A* for a visual explanation).

DESIGN 2. In this design, a paired *t*-test takes advantage of the additional statistical power offered by measuring each individual in both conditions, thus eliminating from the analysis any

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	MASTER LIST				LIST TO MOVE				CONTROL GROUP				EXPTAL GROUP		
2	egg #	mass	$\dot{V}O_2$		egg #	mass	$\dot{V}O_2$		egg #	mass	$\dot{V}O_2$		egg #	mass	$\dot{V}O_2$
3	1	49.6	20.5		1	49.6	20.5		10	54.4	20.8		7	54.9	14.4
4	2	49.2	17.5						4	43.5	18.0		18	49.0	19.1
5	4	43.5	18.0						12	53.4	17.8		2	49.2	17.5
6	6	50.9	14.2		6	50.9	14.2								
7	7	54.9	14.4												
8	8	52.2	15.5		8	52.2	15.5								
9	9	49.4	14.0		9	49.4	14.0								
10	10	54.4	20.8												
11	12	53.4	17.8							= AVERAGE(O3:O14)					
12	14	51.9	19.6		14	51.9	19.6								
13	15	50.3	14.5		15	50.3	14.5								
14	18	49.0	19.1												
15									mean	50.4	18.9		mean	51.0	17.0
16									s.e.m.	3.5	1.0		s.e.m.	1.9	1.4

= STDEV(J3:J14)/SQRT(COUNT(J3:J14))

Fig. 2. Spreadsheet arrangement for creating balanced groups, shown as it would appear in Microsoft Excel in the middle of the process (*eggs 2*, *4*, *7*, *10*, *12*, and *18* have been tentatively assigned to groups, whereas *eggs 1*, *6*, *8*, *9*, *14*, and *15* have yet to be assigned). This example demonstrates the process for creating two groups balanced for the variables of mass and as would be used for *design 1* but can be adapted for any number of groups or variables per group. By trial and error, individual rows of data are cut from the "List to Move" and pasted into either of the two treatment group lists [control or experimental (EXPTAL)] so that the groups are matched as closely as possible for both means and SEs for each measured variable. The "Master List" and List to Move are initially identical, but the Master List provides a reference once rows from the List to Move have been cut and pasted into either the control group or experimental group lists. Cells for means and SEs contain formulas that calculate the new means and SEs for each group each time a row of data for an individual egg is moved into or out of a treatment group list. The "AVERAGE" function is used because it will calculate the mean correctly even when some of the cells in the indicated range are empty. SE [or SD (STDEV)/the square root (SQRT) of *n*] is used rather than the SD because it provides a convenient estimate of whether the two groups are likely to be drawn from the same population (8); as membership in each group grows, the decreasing SE reflects the increasing certainty with which the population mean can be estimated. The goal is to produce groups with similar SEs and in which the ranges of means ± SE for the two groups overlap one another. The formula for SE must be entered by hand as shown; the COUNT function returns the number of filled cells in the indicated range, insuring that the calculation will return the correct value regardless of the number of empty cells in the indicated range.

differences due to an intrinsically high or low $\dot{V}O_2$ in any particular individual. The paired *t*-test computes the difference between the two measurements for each embryo; it then computes the mean of these differences and compares that mean with zero, the mean difference that would result if there were no consistent effect of temperature on $\dot{V}O_2$ (see Fig. 1*B* for a visual explanation). The power of the paired *t*-test is illustrated in Fig. 3.

Assumptions. A student with data in hand is a student who wants a *P* value–and fast! It is important, however, to emphasize to students that they must first determine whether their data meet the assumptions of the test they plan to perform. Students should be in the habit of examining frequency histograms of their data before proceeding with any other kind of analysis.

For the two-sample *t*-test, measurements obtained in each condition should be roughly normally distributed (there should be no obvious outliers, the data should not be bimodally distributed, and the data should not be strongly skewed to high or low values). Students frequently make the mistake of plotting frequency histograms of all the data, which will produce a strikingly bimodal distribution if there is a significant difference between the two treatment groups; they must be sure that the data are plotted separately for each group being compared. In the chicken embryo metabolism experiment, this would mean checking the distributions for the data for control embryos separately from the data for experimental embryos. *t*-Tests that do not assume equal variances are easiest to use; otherwise, any of several tests offered by your statistical software package can be used to test for heterogeneity of variance; note that *P* values of <0.05 indicate that variances are significantly different.

To test the assumptions of the paired *t*-test, data from each individual are reduced to a single value: the difference between the two measurements. The resulting set of differences should be plotted to determine whether they are normally distributed.

If the assumptions of the *t*-test are not met, the easiest solution is to substitute a nonparametric test. For example, the Mann-Whitney *U*-test may be substituted for the two-sample *t*-test or the Wilcoxon matched-pairs test may be substituted for the paired *t*-test; your statistical software package may have slightly different choices. Nonparametric tests are less powerful but make no assumptions about the form of the data; instead of relying on distributions around means, which can be strongly affected by outliers, many nonparametric tests use the relative ranks of the data, which are only minimally affected by outliers. Other solutions, such as transforming the data so that the distribution becomes normal, can also be used (23). It is not possible to include a more thorough discussion of the statistical tests and principles referred to here; readers unfamiliar with basic statistics should consult a textbook or other source, such as *a Biostats Basics* by Gould and Gould (8), which includes references to a website where statistical tests can be performed online.

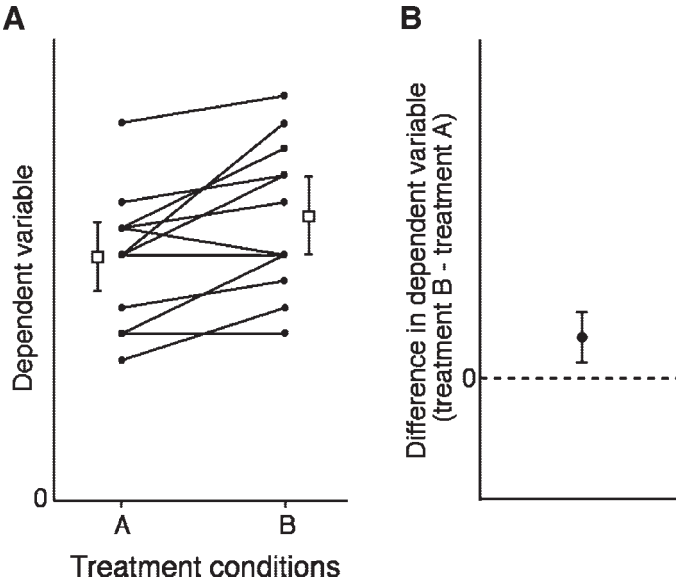

Fig. 3. *A*: measurements taken under two treatment conditions (*conditions A* and *B*). Each pair of points connected by a line represents one subject in the experiment. Because the means for the two set of measurements (□) are very similar and variation among individuals is high, a two-sample *t*-test comparing these means finds no significant difference. Connecting the pairs of points with lines, however, indicates that, within any one subject, measurements in *condition A* are quite consistently lower than those in *condition B*. A paired *t*-test (*B*; different *y*-axis scale) removes differences between individuals by considering only the differences between *conditions A* and *B*. The paired *t*-test computes the mean of these differences and compares the mean with 0, which would be the mean difference if there were no effect of treatment condition on the dependent variable. In the case illustrated here, the two-sample *t*-test found no significant difference (no significant effect of treatment condition), whereas the paired *t*-test found a very significant difference.

General Principles

These are some of the general principles that I hope my students will derive from the exercise of designing a simple experiment. Some of the thought questions included at the end of this article are designed to prompt students to state these principles for themselves. I find that this student-centered method is generally more effective in instilling the principles than simply providing them for students to memorize.

Principle 1. To control for time-related variables, one must always compare experimentally treated animals with controls that, with the exception of the independent variable that is being tested, have experienced identical conditions throughout the experiment. A simple before-after comparison within individuals is not valid because it does not control for extraneous time-dependent factors (see *thought questions 1*, *4*, and *6*)

Principle 2. A control group is a group of animals that receives treatment identical to that of the experimental group except for the independent variable of interest. Similarly, a control treatment is identical to the experimental treatment except for the independent variable of interest (see *thought questions 1*, *2*, and *6–8*).

Principle 3. A well-designed experiment does not necessarily have a "control group." If each individual receives all the treatments in random order, there is no specifically designated control group, but the experiment as a whole is controlled (see *thought question 6*).

Principle 4. Some variables, such as temperature, do not have an obvious control condition because it is not possible to design a treatment lacking this variable. In such cases, a controlled experiment compares two treatments that are identical except for the variable of interest (e.g., temperature) (see *thought questions 6* and 7).

Principle 5. The goal of random assignment to treatment groups is the creation of balanced groups. However, when sample sizes are small and/or there is large variation among individuals, actively balancing groups before the experiment starts is a more reliable method than random assignment for creating balanced groups (see *thought questions 2* and *4*).

Principle 6. Experiments that can be analyzed with a paired statistical test are powerful because they eliminate baseline differences among individuals (see *thought question 8*).

Thought Questions

These questions are designed to help students discover some of the general principles of experimental design and analysis in an active way (see *Classroom Implementation* for specific suggestions on using these questions to supplement your classroom discussion). Depending on the way in which you use these questions, you may want to change the order of the questions or omit some of them. Suggested answers are shown in italics.

Question 1. The $\dot{V}O_2$ of 10 frogs was measured shortly after the frogs had been captured in the wild and brought into the laboratory. An insecticide was then added to the aquarium water of all 10 frogs at a dose similar to what the frogs might experience in their natural habitat. After 4 wk of insecticide treatment, the $\dot{V}O_2$ of the 10 frogs was measured again.

A. The results of the experiment are shown in Fig. 4*A*. What statistical test would you use to analyze these data? Why? *Answer: because the same individuals were measured in both conditions, a paired t-test would be the appropriate test. Students may recognize, and the instructor may want to discuss, that this is an inadequately controlled experiment, since there are no controls for time-dependent variables.*

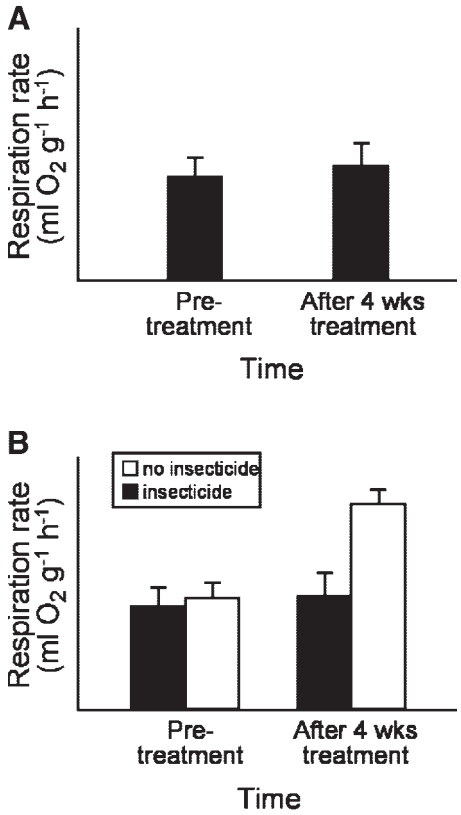

Fig. 4. $\dot{V}O_2$ of frogs before and after 4 wk of treatment with insecticide.

B. From the results shown in Fig. 4*A* alone, what would you conclude about the effect of this insecticide on the $\dot{V}O_2$ of the frogs? *Answer: results suggest that the insecticide has no effect on the $\dot{V}O_2$ of the frogs.*

C. The results reported in *question 1A* were in fact only some of the results obtained in this experiment. Actually, 20 frogs were captured for the frog study. The $\dot{V}O_2$ of all 20 frogs was measured shortly after the frogs had been captured. However, the aquarium water for the second group of 10 frogs was not treated with the insecticide. The $\dot{V}O_2$ of this control group of 10 frogs was also remeasured after 4 wk. The complete set of results obtained in this experiment is shown in Fig. 4*B*. Assuming that no mistakes were made in the experiment or in the collection of the data, and that all of the frogs were in equally good health when captured, propose a biological explanation for the results shown in Fig. 4*B*. *Answer: the control animals demonstrate that over time in captivity, $\dot{V}O_2$ increases. A possible explanation is that the animals may have been affected by the stress of capture and the new captive environment when they were first brought into the laboratory and that as they habituated to this new environment, this response abated and $\dot{V}O_2$ increased. However, the $\dot{V}O_2$ of insecticide-treated frogs did not change over time; possible explanations are that the insecticide induced a continuous stress state or that the insecticide directly inhibited $\dot{V}O_2$ in some other way. Students with more detailed knowledge of the molecular pathways in cellular respiration could be asked to follow up their hypothesis by proposing particular molecular targets of the insecticide.*

D. Based on all of the data collected in this experiment, what do you conclude about the effect of the insecticide on the $\dot{V}O_2$ of the frogs? *Answer: the insecticide reduces the $\dot{V}O_2$ of the frogs.*

E. Which sets of measurements did you compare in order to reach your conclusion? *Answer: the 4-wk measurements for insecticide-treated frogs should be compared with the 4-wk measurements for the control frogs.*

F. What statistical test would you use to make this comparison? Why? *Answer: a two-sample t-test should be used because the two treatment groups consist of different individuals.*

G. What kinds of variables were unaccounted for in the first data set (Fig. 4*A*)? *Answer: the effects of time-dependent variables, such as habituation to the new captive environment and learning.*

H. If you were the CEO of the company that manufactures the insecticide used in this experiment, which results would you want to publish? *Answer: although they would misrepresent the experiment, the company might want to publish only the results shown in Fig. 4a because it casts their product in a more favorable light.*

I. If you were called to testify as an expert witness against a company that claimed its products were safe on the basis of the results shown in Fig. 4*A*, what would you tell the jury? Your explanation must make your point clear to 12 nonscientists! Try your explanation out on the rest of the class, and vote for the most effective explanation.

J. Write a general statement or rule for how one should set up and analyze the results of a "before-after" (also known as a "pretest-posttest") experiment such as this. Write this rule in

your notebook. *Answer: to control for time-related variables, one must always compare experimentally treated animals with controls that, with the exception of the independent variable under investigation, have experienced identical conditions throughout the experiment (see general principles 1 and 2).*

Question 2. During the winter, some small mammals are able to reduce their body temperature and metabolic rate for a few hours each day in a hibernation-like state known as daily torpor. A researcher conducted the following experiment to determine whether cortisol, a hormone involved in the steroid stress response, would increase the use of daily torpor by hamsters. Sixteen hamsters were implanted with temperature sensors that provided a continuous record of body temperature. The animals were then divided randomly into two groups, each containing eight animals. During the first week of the experiment, each animal's body temperature records were used to determine the number of torpor bouts that occurred over that week for each hamster. During the second week of the experiment, one group was given a daily dose of cortisol plus vehicle (cyclodextrin) in the drinking water; the other group consumed drinking water that contained only cyclodextrin. The data from this experiment are shown in Fig. 5.

A. What is meant by "vehicle" in this experiment? Why is it important that both groups of hamsters receive vehicle in their drinking water during *week 2* of the experiment? *Answer: vehicle refers to the putatively inactive substance(s) added along with the test substance to allow delivery of the test substance to a test subject. Vehicles may be solvents (such as a saline solution), may enhance solubility, or may increase absorption into the body. In this example, cyclodextrins are ring-shaped carbohydrates that can be complexed with insoluble hormones such as cortisol to make them soluble in water. To control for any effects of the vehicle itself, vehicle must be given to the controls as well.*

B. Based on the results of this experiment, what would you conclude about the effect of cortisol on daily torpor in hamsters? *Answer: it appears that cortisol has no effect because it did not change the number of torpor bouts in either treatment group.*

C. In many experiments, animals are divided randomly into two treatment groups; one group is then given the experimental treatment and the other group serves as the control, and a single set of measurements is made on both groups. Which of the measurements shown in Fig. 5 correspond to the data that would be collected in such an experiment? Circle them in Fig. 5. Under what conditions does such an experiment produce meaningful results? *Answer: the data from a single measurement experiment correspond to the two measurements labeled "during treatment." Such an experiment produces meaningful results when there are no significant differences in the dependent variable between the groups before treatment begins and when the two groups are as similar to each other as possible so that the differences between the two groups after treatment can be attributed to the independent variable (cortisol) and not to other differences between the groups.*

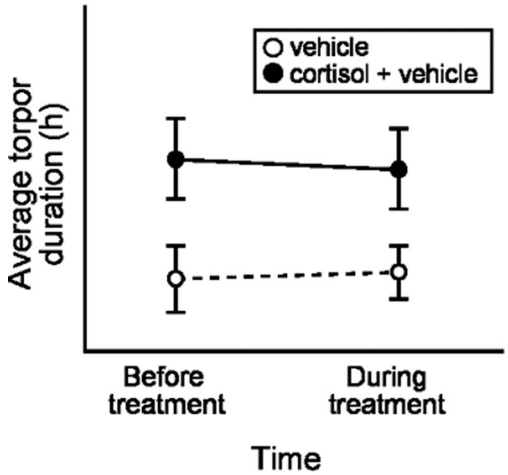

Fig. 5. Average torpor duration over 7 days measured before and during treatment with cortisol in drinking water.

D. If you had applied the general rule that you wrote in response to *question 1J*, what would you have concluded about the effect of cortisol on daily torpor in this experiment? What additional rule must the data follow to prevent you from drawing incorrect conclusions in an experiment like this? Write this rule in your notebook. *Answer: if one simply compared the control and treatment groups during the week of treatment, one would conclude that cortisol increases the use of torpor in hamsters. Additional principle: treatment groups must be balanced so that any differences in response can be attributed to the independent variable being tested.*

E. What could account for the differences in the frequency of daily torpor observed during *week 1* of the experiment? List as many possibilities as you can. *Answer: individual animals can vary tremendously in their use of daily torpor; in a variety of mammalian species that display daily torpor, some individuals are inherently more "torpor prone" than others (18); in rodents displaying daily torpor on a seasonal basis (typically in winter), some individuals are photononresponders that fail to adopt the winter phenotype, including the use of spontaneous daily torpor, even when they are exposed to the short days typical of winter (17). In some species, there may be sex differences as well (11). If random assignment had resulted in groups with different proportions of torpor-prone and torpor-resistant animals, the data obtained before treatment would likely differ significantly. The smaller the sample size, the more likely that random assignment of individuals to treatment groups will result in unintended, significant differences between groups.*

F. Now consider a situation in which the two randomly selected treatment groups each contain 50 rather than 8 hamsters. Is the likelihood of creating groups with significantly different torpor patterns at the beginning of the experiment greater or less in the 100-animal experiment than in the 16-animal experiment? *Answer: this is much less likely in the 100-animal experiment. Larger sample sizes reduce the probability that unbalanced groups will be created through random assignment.*

G. If the experiment design were changed so that torpor in each animal is measured for 1 wk with vehicle treatment and for another week with vehicle plus cortisol, how should the experiment be designed so that it is adequately controlled? *Answer: an alternative design is to subject each individual to both treatments, with one-half of the animals receiving vehicle only first and the other half of the animals receiving vehicle plus cortisol first.*

H. What statistical test would you use to analyze the results of the experiment described in *question 2G* above, and what exactly would you compare? *Answer: a paired t-test should be used to compare all measurements in one treatment (vehicle only) with all measurements in the other treatment (vehicle plus cortisol).*

I. For some kinds of independent variables, an experiment in which each animal receives both treatments is highly effective; for other kinds of independent variables, it is less effective. What kinds of independent variables fall into each of these categories (effective or less effective)? Give as many examples of each type of independent variable as you can. *Answer: independent variables without long-lasting effects (e.g., acute exposure to a particular ambient temperature or short-acting hormones) work well with this kind of design, whereas those with long-lasting effects (e.g., some steroid and thyroid hormones, lipopolysaccharide treatment, addictive drug treatment, or any treatment to which animals habituate rapidly) are less suitable. In the latter case, although the experiment is technically controlled, the aftereffects of the previous treatment may reduce the power of the experiment to reveal any biological effects because the effect of the paired analysis is to average the responses of all experimental subjects.*

Question 3. When animals are divided into treatment groups for experiments, the goal is to produce balanced groups. For each group of animals described below, state whether you think using random assignment to create two groups of equal size would be successful in producing balanced groups. Assume that you will use all of the animals in each example (e.g., if you start with 20 animals, you will create 2 groups of 10 animals). Briefly explain your reasoning.

A. Twenty inbred mice of the same age and sex. *Answer: the experiment is lighly likely to be successful because individuals are extremely similar, not only in genetic makeup but also in age and sex.*

B. Twenty inbred mice of unknown age. *Answer: the experiment is likely to be successful, but the risk of unintentionally creating unbalanced groups is increased because age, and possibly sex, are not taken into account. If the ages and/or sex are unknown but equal for all mice, this risk is reduced. Body mass may serve as a useful proxy for age, especially if the animals are relatively young.*

C. Twenty outbred captive mice. *Answer: the experiment is quite likely to be successful, but slightly more risky than that for inbred mice because there is greater genetic diversity.*

D. Twenty recently captured wild mice. *Answer: the experiment is much less likely to be successful, since individuals can vary considerably in genetic makeup, age, and experience.*

E. One hundred recently captured wild mice. *Answer: increasing the sample size greatly decreases the risk of creating unbalanced groups by random assignment. This experiment is much more likely to be successful than that in D.*

F. Based on your answers to the above questions, write some general rules for identifying situations in which random assignment is a good choice for creating experimental treatment groups and situations in which random assignment is unlikely to be a good choice for creating experimental treatment groups. Write these rules in your notebook. *Answer: random assignment is most likely to be a good choice for genetically similar animals (e.g., inbred laboratory strains) of known age and sex and for large sample sizes. This method is less desirable for wild animals, small sample sizes, or any situation in which there is considerable individual variation in the trait(s) being measured.*

Question 4. Figure 6 shows the results of a study in which $\dot{V}O_2$ was first measured in all animals at 30°C (time = 0 min). These baseline $\dot{V}O_2$ were used to separate the animals into two groups. One of the groups (experimental) was then subjected to a new temperature (20°C), while the other group (control) remained at the original temperature. After 90 min, when the animals had equilibrated at the new temperatures, the $\dot{V}O_2$ of all animals was measured again. (Note that although error bars would normally be included in such a graph, they have been omitted for the sake of clarity.)

A. The two-headed arrows on the graph represent four possible statistical comparisons. For each comparison (*A–D*), name the specific statistical test you would use if this was the *only* comparison you were making. For each comparison (*A–D*), if you were to find a significant difference in this particular comparison, which independent variable(s) would be responsible for this difference? *Answer: for comparison A, a paired t-test shows differences due to temperature plus all time-dependent extraneous variables such as habituation, learning, prior handling, etc. For comparison B, an unpaired t-test shows differences due to temperature alone. For comparison C, a paired t-test shows differences due to time-dependent extraneous variables alone. For comparison D, a two-sample t-test shows baseline differences between groups due to individual differences in* $\dot{V}O_2$.

B. It is best to limit the number of statistical comparisons that are made in a single experiment. Fortunately, only one comparison shows the effect of temperature alone. Which comparison is this? *Answer: comparison B.*

C. Why is it important to balance the two treatment groups after the baseline measurement is made? *Answer: if groups are not balanced initially, at least some of the difference in comparison D may be due to individual differences rather than to the independent variable, temperature.*

Question 5. Figure 7 shows the results of an experiment testing the effect of a mild sedative on activity in a nocturnal rodent. Baseline activity data were recorded at *time 0* and used to create balanced treatment groups. Activity was then recorded again after each group had received its treatment: the

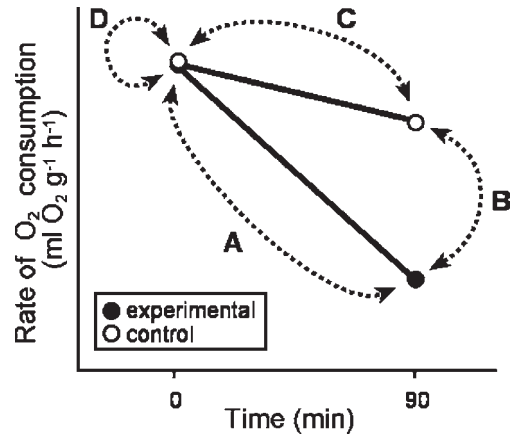

Fig. 6. $\dot{V}O_2$ of animals at two temperatures. Controls were measured at 38°C at both time points, whereas experimental animals were measured first at 38°C (time = 0 min) and then at 23°C (time = 90 min). *Comparisons A–D* are shown.

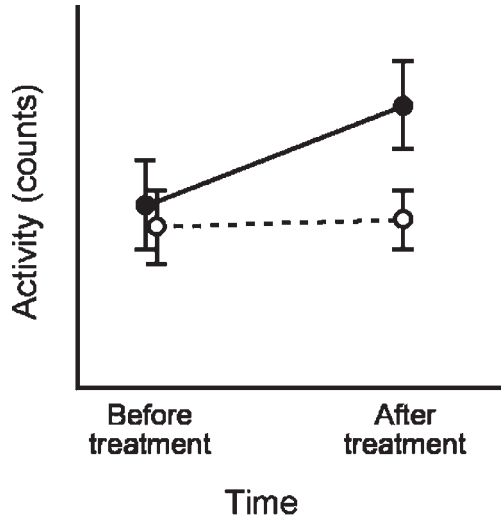

Fig. 7. Activity (measured as numbers of revolutions of a running wheel placed in the cage) of animals at two time points. The first set of measurements was made before treatment, and the second set was made after one of the groups had been treated with a mild sedative plus vehicle and the other group had been treated with vehicle alone.

experimental group was treated with sedative plus vehicle, whereas controls were treated with vehicle only.

A. Indicate on the graph which group is the experimental group and which is the control group. What do you conclude about the effects of the sedative from these data? *Answer: most students will assume that the group with little change in activity represents the control group. With this assumption, one would conclude that the sedative increases activity.*

B. Although most people assume that the group with the least change over time is the control group, this is not necessarily true. The reverse might be true, for example, if the activity of the nocturnal rodents in this experiment was first measured during the day (when nocturnal rodents are sleeping) and then measured again during the night, after treatment with the mild sedative. Because the animals are nocturnal, their activity at night with the sedative is still higher than that during the day, when they are completely inactive with or without the sedative. Switch the designations of "control" and "experimental" that you indicated in *question 5A* above. Now restate the conclusion you would draw from this experiment. *Answer: the sedative decreases activity in this species.*

C. Propose another explanation for why the activity of the controls might increase after treatment with vehicle while the activity of the animals treated with vehicle plus sedative remains the same. *Answer: one hypothesis is that the vehicle itself may have the unintended effect of increasing activity, but, when given in combination with the sedative, it results in no change in activity. Note, however, that this alternate explanation does not alter the conclusion reached by comparing the activity of the two groups after treatment, namely, that the effect of the sedative is to reduce activity in this species.*

D. What general points (or principles) are illustrated by these data? *Answer: one must always compare control with experimental treatments (rather than experimental animals with themselves) in this experimental design. In addition, controls do not always stay the same over the course of an experiment; this is the very reason for needing a control group (as in design 1) or a control for treatment order (as in design 2).*

Question 6. In a recent study, researchers tested the cardiovascular response of human subjects to immersing their faces in water. The heart rate of each subject was measured after the subject had been resting quietly for 1 min. The subject then immersed his or her face in cold water, and the heart rate was measured again.

A. Explain what can be learned from this experiment. Explain what *cannot* be learned from this experiment. *Answer: this experiment tells us the combined effects of time since the first heart rate measurement plus the effect of immersion in cold water. It cannot tell us about the effects of cold water immersion alone.*

B. If you could redesign this experiment, how would you change it so that it is properly controlled? In your explanation, include the name of the statistical test you would use to analyze the data, and specify exactly which data you would compare with this statistical test. *Answer: measure the resting and immersion heart rate in all subjects, but measure resting rates first in one-half the subjects and immersion rates first in the other subjects. Use a paired t-test to compare all resting heart rates with all immersion heart rates.*

C. In the redesigned experiment you described in *question 6B*, which group is the control group? *Answer: neither group is the control group. The experiment is controlled by the random order of treatments rather than by comparison with a control group. The resting treatment may be considered a control treatment; both groups experience the control condition but in different orders.*

Question 7. What is a control? Write a general definition that could apply to any experiment in which a group of subjects is designated as a control group. How would you describe the control condition (or treatment) in an experiment in which all subjects experience the control condition but in different orders? *Answer: a control group is a group of animals that receives treatment identical to that of the experimental group except that they are not exposed to the independent variable of interest (general principle 2). Similarly, a control condition or treatment is identical to the experimental treatment except that the subjects are not exposed to the independent variable of interest. Note that what constitutes the control condition for an omnipresent, continuous variable such as ambient temperature is not always obvious. The experiment may simply be comparing what happens at two temperatures, but the rules for setting up and analyzing the results of the experiment are the same.*

Question 8. Does every well-designed experiment have a group that is designated the "control?" Describe a controlled experimental design in which no single group is the control group. *Answer: see experimental design 2 (Fig. 1). See also the final note in the suggested answer to thought question 7.*

ACKNOWLEDGMENTS

I thank Dee Silverthorn for the astute and helpful suggestions; Steven Wang of the Swarthmore College Department of Mathematics and Statistics for statistical consulting; Itzick Vatnick for data on which *thought question 1* are based; Swarthmore College and a Lang Fellowship for sabbatical funding; Fritz Geiser at the University of New England for a fine intellectual home during the sabbatical when this manuscript was written; my colleagues in the Biology Department at Swarthmore College for stimulating discussions regarding pedagogy; and the many generations of Animal Physiology students who have helped to shape my teaching.

REFERENCES

1. **Brown JH, West GB.** *Scaling in Biology*. New York: Oxford Univ. Press, 2000.
2. **Calder WA.** *Size, Function and Life History*. Cambridge, MA: Harvard Univ. Press, 1984.
3. **Donovan MS, Bransford JD** (editors). *How Students Learn: History, Mathematics and Science in the Classroom*. Washington, DC: National Academies, 2005.
4. **Festing MF.** Principles: the need for better experimental design. *Trends Pharmacol Sci* 24: 341–345, 2003.
5. **Festing MF.** Reduction of animal use: experimental design and quality of experiments. *Lab Anim* 28: 212–221, 1994.
6. **Gandhi M, Aweeka F, Greenblatt R, Blaschke T.** Sex differences in pharmacokinetics and pharmacodynamics. *Annu Rev Pharmacol Toxicol* 44: 499–523, 2004.
7. **Gotelli NJ, Ellison AM.** *A Primer of Ecological Statistics*. Sunderland, MA: Sinauer, 2004.
8. **Gould JL, Gould GF.** *Biostats Basics*. New York: Freeman, 2002.
9. **Hiebert SM.** The strong inference protocol: not just for grant proposals. *Adv Physiol Educ* 31: 93–96, 2007.
10. **Hiebert SM, Noveral J.** Are chicken embryos endotherms or ectotherms? A laboratory exercise integrating concepts in thermoregulation and metabolism. *Adv Physiol Educ* 31: 97–109, 2007.
11. **Hiebert SM, Wingfield JC, Ramenofsky M, Deni L, Gräfin zu Elz A.** Sex differences in the response of torpor to exogenous corticosterone during the onset of the migratory season in rufous hummingbirds. In: *Life in the Cold: Evolution, Mechanisms, Adaptation and Application. Twelfth International Hibernation Symposium*, edited by Barnes BM and Carey HV. Fairbanks, AK: Univ. of Alaska Fairbanks, 2004, p. 221–230.
12. **Hodes GE, Shors TJ.** Distinctive stress effects on learning during puberty. *Hormones Behav* 48: 163–171, 2005.
13. **Institute of Laboratory Animal Resources, National Research Council.** *Guide for the Care and Use of Laboratory Animals*. Washington, DC: National Academy, 1996.
14. **Kaiser J.** Gender in the pharmacy: does it matter? *Science* 308: 1572–1574, 2005.
15. **Mendelsohn ME, Karas RH.** Molecular and cellular basis of cardiovascular gender differences. *Science* 308: 1583–1587, 2005.
16. **Nowak R.** Problems in clinical trials go far beyond misconduct. *Science* 264: 1538–1541, 1994.
17. **Puchalski W, Lynch GR.** Circadian characteristics of Djungarian hamsters: effects of photoperiodic pretreatment and artificial selection. *Am J Physiol Regul Integr Comp Physiol* 261: R670–R676, 1991.
18. **Ruf T, Heldmaier G.** The impact of daily torpor on energy requirements in the Djungarian hamster, *Phodopus sungorus*. *Physiol Zool* 65: 994–1010, 1992.
19. **Schmidt-Nielsen K.** *Scaling: Why Is Animal Size So Important*? New York: Cambridge Univ. Press, 1984.
20. **Shiland TW.** Constructivism: the implications for laboratory work. *J Chem Ed* 76: 107–109, 1999.
21. **Singer SR, Hilton ML, Schweingruber HA** (editors). *America's Lab Report: Investigations in High School Science*. Washington, DC: National Academies, 2005. http://newton.nap.edu/books/0309096715/html [16 November 2006].
22. **Snedecor GW, Cochran WG.** *Statistical Methods* (6th ed.). Ames, IA: Iowa State Univ. Press, 1967, p. 110.
23. **Sokal RR, Rohlf FJ.** *Biometry*. New York: Freeman, 1995.
24. **Spearow JL, Doemeny P, Sera R, Leffler R, Barkley M.** Genetic variation in susceptibility to endocrine disruption by estrogen in mice. *Science* 285: 1259–1261, 1999.
25. **Spicer JI, Gaston KJ.** *Physiological Diversity and Its Ecological Implications*. London: Blackwell Science, 1999.
26. **Webb DR.** Thermal tolerance of avian embryos: a review. *Condor* 89: 874–898, 1987.
27. **Wizemann TM, Pardue ML.** *Exploring the Biological Contributions to Human Health: Does Sex Matter?* Washington, DC: Board on Health Sciences Policy, Institute of Medicine, 2001.
28. **Zolman JF.** Teaching experimental design to biologists. *Adv Physiol Educ* 22: 111–118, 1999.