CONTENTS

Introduction vii

1. Conducting Meaningful Experiments: Prerequisites and Purposes 1

2. Causal Inferences and the Strange (but Fictitious) Case of Mrs. Smith 7

3. Introduction to the Design of Meaningful Experiments via the Continuing Adventures of Dr. Jones 11

4. Why Poorly Designed Experiments Are Inadequate for Making Complex Inferences: The Single-Group Pretest-Posttest Design (or the Unfortunate Conclusion of Dr. Jones’s Foray into the World of Science) 17

5. Enter the Control Group and a deus ex machina 23

6. The Design of Single Factor, Between-Subjects Experiments 33

7. Factorial Designs 43

8. Repeated Measures, Within-Subjects, and Longitudinal Designs 55

9. Ensuring Sufficient Statistical Power 65

10. Conducting an Experiment 73

11. External Validity (Generalizability) 77

12. Three Additional Experimental Paradigms: Single-Case, Program Evaluation (Including Natural Experiments), and Quality Improvement Research 81

13. Experimental Bias 87

A Closing Word 95
INTRODUCTION

The purpose of this manual is to supplement the use of The Design and Conduct of Meaningful Experiments Involving Human Participants in the teaching of experimental design concepts. It has been written to serve as a collaborative endeavor between the text's author and the instructor of any course that employs the book.

Hopefully, therefore, this manual will save you, the instructor, a bit of time by making suggestions for (a) class activities, (b) discussion topics, and (c) sample items that are designed to facilitate student understanding of the concepts presented in the text.

A textbook author and an instructor each have certain advantages over one another as well as differing constraints. You interact face to face or online with the students and can therefore both respond to their questions regarding concepts they don't understand and supply supplementary instruction, activities, and assignments as you perceive the need. I, on the other hand, having never seen your students (or even know your or their parent disciplines) can only anticipate their current needs and their potential future needs as aspiring scientists in all disciplines involving experimentation with human participants. To compensate, I have had the time to craft as careful and thorough explanations of this content in the book itself as I was capable of within my rather limited writing style. Since I conceive of this supplementary manual as a means of our working together toward our common objectives, please feel free to contact me at: bausell@gmail.com if there is anything I can clarify from my end.

The manual is organized by book chapters. For each chapter I have attempted to first summarize what I consider the most important points for students to internalize. I then present a limited number of possible activities followed by 10 or so thought questions at the end of each chapter that might be posed to students after they have had time to read the appropriate material.

SAMPLE ITEMS

I have taken somewhat of a risk with the format and the type of questions I have posed, which I hope will not be overly irritating to students. For the most part I purposefully avoided factual

INTRODUCTION

The purpose of this manual is to supplement the use of The Design and Conduct of Meaningful Experiments Involving Human Participants in the teaching of experimental design concepts. It has been written to serve as a collaborative endeavor between the text's author and the instructor of any course that employs the book.

Hopefully, therefore, this manual will save you, the instructor, a bit of time by making suggestions for (a) class activities, (b) discussion topics, and (c) sample items that are designed to facilitate student understanding of the concepts presented in the text.

A textbook author and an instructor each have certain advantages over one another as well as differing constraints. You interact face to face or online with the students and can therefore both respond to their questions regarding concepts they don't understand and supply supplementary instruction, activities, and assignments as you perceive the need. I, on the other hand, having never seen your students (or even know your or their parent disciplines) can only anticipate their current needs and their potential future needs as aspiring scientists in all disciplines involving experimentation with human participants. To compensate, I have had the time to craft as careful and thorough explanations of this content in the book itself as I was capable of within my rather limited writing style. Since I conceive of this supplementary manual as a means of our working together toward our common objectives, please feel free to contact me at: bausell@gmail.com if there is anything I can clarify from my end.

The manual is organized by book chapters. For each chapter I have attempted to first summarize what I consider the most important points for students to internalize. I then present a limited number of possible activities followed by 10 or so thought questions at the end of each chapter that might be posed to students after they have had time to read the appropriate material.

SAMPLE ITEMS

I have taken somewhat of a risk with the format and the type of questions I have posed, which I hope will not be overly irritating to students. For the most part I purposefully avoided factual
items in favor of questions that required a degree of understanding of the concepts involved. For multiple-choice questions I felt that the most effective, non-redundant means of doing this was to usually permit more than one correct response, although I am well aware that this does not constitute good psychometric practice. The items, however, were not designed to produce “reliable” or “valid” summative scores, but rather to be a bit more difficult and thought provoking than students will expect to account on a timed, graded quiz or exam. I would advise you not to use them in your course’s grading system. Instead, I think they would be better used as questions for discussion purposes or mock quizzes, the answers to which can then be challenged since many of the “correct” answers supplied are situational specific and a few of the “incorrect” answers could be acceptable under certain very specific conditions. For this reason, rather than simply presenting a scoring key, I have presented brief rationales for why each response is either correct or incorrect based on what is presented in the text.

EXAMPLES

The targeted interdisciplinary nature of the book makes it difficult to identify specific examples that will resonate with all students. If you are blessed with a TA or an RA, I would suggest sending them to the literature to supplement some of my examples. If you don’t have such help, hopefully my choices will suffice.

READING AND CRITIQUING RESEARCH REPORTS

There is no magic number for how many experimental articles students should read and critique, but in my opinion it should be as many as the market will allow, and the task should be distributed throughout the semester. It might also be helpful if you assigned everyone an article of your choice for the first assignment so that it can be discussed in class and the students can learn from one another.

Certainly no student wants to write, and no instructor wants to read, lengthy, wordy descriptions of the myriad studies that will result from this activity, hence you might want to simply allow the student to cut and paste the article’s abstract, perhaps answer a few standard questions, and then discuss the strengths (and especially the weaknesses of the study itself and the methodology employed). Examples of some standard, short-answer questions that might be used are as follows:

1. Was the purpose, hypothesis, or objective of the experiment clearly stated? Since few if any of the selected articles will have the five-point, detailed hypotheses I have advocated for pedagogical purposes in Chapter Three, another component of the exercise could be to have the students compose one as part of their critique.
2. What was the design employed (e.g., a single-group pretest-posttest)?
3. Were the design and procedures employed adequate to achieve the stated purpose?
4. Did the author provide sufficient detail for you to evaluate the study? If not, what would you have liked to know that was not revealed?
5. Did the discussion match the results and the experimental conditions (e.g., was the discussion hyped or measured)?
6. Was the experiment meaningful in your opinion?

Now of course there is much more to be critiqued about a research report than this, such as the adequacy of the (a) literature review, (b) sampling, (c) measurement, (d) statistical analysis, and so forth, but generally beginning students don’t have the expertise to evaluate these factors and the critiques can correspondingly become mechanistic (and boring) in nature. Obviously, if you provide instruction in these topics, or the students have had prerequisite courses in them, then by all means include them.

STUDENTS' DESIGNING THEIR OWN EXPERIMENTS

The process of students designing their own experiment, preferably one which they believe would make a contribution to their discipline, can be both a rewarding endeavor in its own right and an effective teaching mechanism. After an initial dry run, this should probably be a semester-long project for some students while others may have a head start by being at or near the dissertation phase of their programs.

Regardless, at some intermediary point of this process they should present an outline of their proposed study and its design to you for approval and suggestions, after which they can present it to the class or to roundtable groups of students (who by then will hopefully be capable of critiquing the proposal and offering suggestions). The end product can be a full written proposal (or a truncated one) that you would formally evaluate. If time permits, the proposal could be presented to the class as a whole or perhaps via a poster session to which other faculty and students are invited.

A WORD ABOUT STUDENTS' CONDUCTING THEIR OWN EXPERIMENTS

When I first started teaching research methods it was possible for students to actually conduct their own group projects using other students or healthy cooperating adults (e.g., health professionals) as subjects via a blanket local IRB exemption for research conducted for educational purposes. Those days are long gone, of course, so chances are that this won’t be an option. (Nor is it a great loss anyway because the studies actually conducted were quite rudimentary.)

INSTRUCTOR DESIGNED EXPERIMENTS USING THE CLASS AS PARTICIPANTS

However, it may be possible for you to design small-scale experiments using your class as participants if responses are anonymous and everyone has the right of refusal. (It might still be a
INTRODUCTION

good idea to check with your IRB, the argument being that the results will not be published and the activity is used strictly for educational purposes. If the IRB is unmoved, perhaps the process of securing approval would constitute a teaching moment.)

Not knowing your discipline makes it difficult for me to suggest such projects, but they are quite feasible in many disciplines for which well-known (and reliable) effects exist (such as education and psychology). The most powerful and reliable intervention for increasing learning, for example, is time on task, so one possibility would be to do the following:

1. Select a prose passage, dense with facts, and written on a topic that students can be assumed to know very little about (Wikipedia is good source since it has such entries on every topic known to humankind). An example of a qualifying topic might be the history of parliamentary governments during the Ottoman Empire. Obviously, you should feel free to select a passage of your own. (Naturally, you’ll have to relax the book’s meaningfulness criteria since these are exercises and not real experiments.) Also naturally, after extensive work you might even be able to come up with a somewhat more interesting topic. In any case, the following “fascinating” description is a direct quote from Wikipedia (http://en.wikipedia.org/wiki/Grand_National_Assembly_of_Turkey):

There were two parliamentary governments during the Ottoman period in what is now Turkey. The first constitutional era lasted for only 2 years, elections being held only twice. After the first elections, there were a number of criticisms of the government due to the Russo-Turkish War, 1877–1878 by the representatives, and the assembly was dissolved and an election called on June 28, 1877. The second assembly was also dissolved by the sultan Abdülhamid II on February 14, 1878, the result being the return of absolute monarchy with Abdülhamid in power and the suspension of the Ottoman constitution of 1876, which had come with the democratic reforms resulting in the first constitutional era.

The second constitutional era is considered to have begun on July 23, 1908. The constitution that was written for the first parliament included control of the sultan on the public and was removed during 1909, 1912, 1914, and 1916, in a session known as the “declaration of freedom.” Most of the modern parliamentary rights that were not granted in the first constitution were granted, such as the abolition of the right of the sultan to deport citizens that were claimed to have committed harmful activities, the establishment of a free press, and a ban on censorship. Freedom to hold meetings and establish political parties was recognized, and the government was held responsible to the assembly, not to the sultan.

During the two constitutional eras of the Ottoman Empire, the Ottoman parliament was called the General Assembly of the Ottoman Empire and was bicameral. The upper house was the Senate of the Ottoman Empire, the members of which were selected by the sultan. The role of the Grand Vizier, the centuries-old top ministerial office in the empire, transformed in line with other European states into one identical to the office of a Prime Minister, as well as that of the Speaker of the Senate. The lower chamber of the General Assembly was the Chamber of Deputies of the Ottoman Empire, the members of which were elected by the general public.

2. Construct a 15- to 20-item short answer test from this material (e.g., How many parliamentary governments occurred during the Ottoman period? How long did the first one last?). Using free recall of all the facts retained from the passage would be a more sensitive outcome but it would also take much longer to grade. Multiple-choice items are generally not sensitive enough for research purposes.
3. Decide on an intervention. (Obviously we're going about designing this study backwards, but the students won't be privy to this fact, and it is not destined to go down in the annals of great educational experiments of the 21st century.) Possibilities here include reading the passage vs. not reading the passage, studying the passage for 90 seconds vs. 5 minutes, providing one group with cues to what will be on the test (such as by underlining key sentences that contain the answers to the test items) and the other group with no cues, and so forth.

4. Construct a brief, informed consent document for students to sign, not so much for IRB purposes (assuming you aren't required to go through the IRB) but as a teaching tool.

5. Randomly assign students to one of the two groups, designate one student as your RA, separate the two groups (if there is an empty classroom close by; if not, simply use written instructions), administer the instructions to one group and have your "RA" do the same for the other group reading from a prepared script, conduct the experiment, and administer the test.

6. Bring the students back together and give them an answer sheet and have them score each other's tests. Calculate group means and present the results on the board. Preferably, use an online calculator (such as http://graphpad.com/quickcalcs/ttest1.cfm, as discussed later in this manual under one of the Chapter Five activities) and actually analyze the results for the entire class. Discuss the design, what the hypothesis should be, why they think the results came out the way they did, and so forth.

A WORD ON THE LACK OF WORDS ON REGULATORY MATTERS

As mentioned in Chapter Three, so much variability exists with respect to interdisciplinary and inter-institutional differences in IRB requirements and regulatory requirements that this book does not cover these issues in detail. Therefore, this may be an area on which you want to spend some extra time. If your institution has online training in IRB issues or experimental ethics (which the book does cover) I would suggest a course requirement being completion of that training. (All the IRB training courses with which I have had experience included scenarios followed by quizzes, which can be helpful to students.)

ADDITIONAL READINGS

I haven't overburdened the manual with additional readings. The few that I have mentioned are mainly free and I've provided links that students can access.

EXPERIMENTAL PRINCIPLES

As far as I know, a unique aspect of this book (other than another, much briefer book I wrote over 20 years ago) is the presentation of 25 experimental principles that are simply too important to bury in the text. (I consider them important enough, in fact, to actually be called dicta,
or perhaps commandments if that term hadn’t been taken.) Most of these are design oriented, but some involve scientific mores and acculturation issues. If I’ve left out any that you consider important, by all means feel free to add them to the list.

EXPERIMENTAL ARTIFACTS

Also bolded and separated from the text itself are what I label experimental artifacts; others before me have called them threats to internal, external validity, and statistical conclusion validity (also alternative explanations for experimental results or experimental confounds). There are 18 of these presented, many of which (but not all) have been borrowed from either


I believe that all 18 should be learned and internalized to the maximum degree possible by students who want or need to understand experimental research.
This introductory chapter is heavy on acculturation issues and is designed to introduce students to the experimental process. Obviously, a very small percentage of any class (unless it is a doctoral seminar of some sort) is likely to go on to become practicing scientists, but I don’t believe it is necessary to say this to students. Perhaps the assumption that students are in the class in preparation to becoming scientists (and teaching the content from this perspective) will make the acculturation principles more relevant (and hopefully will help them to become internalized).

DEFINITION OF PROSPECTIVE EXPERIMENTATION

A prospective experiment (which basically constitutes the subject matter of this text) was defined as:

A prospective activity in which (a) an intervention is administered, (b) outcome data are collected, and (c) the results are formally analyzed to ascertain the effects of the intervention upon that outcome.

Note that implicit in this definition is that a comparison of some sort is involved.

RELATED ACTIVITY

There is nothing sacrosanct about this definition and most likely it can be improved, but students should probably be given a definition of some sort along with the clarification that an
intervention is very broadly defined as anything purposefully done within the experimental setting designed to change participants’ outcomes. As such, therefore, the generic purpose of an experiment is to generate a valid, causal inference.

Formulating a definition of experimental research: Prior to reading the chapter, ask the class to write a single-sentence definition of a prospective experiment. (As mentioned in the introduction, these activities and questions can be administered individually or to groups.)

Critiquing a definition of experimental research: After reading the chapter (and perhaps prior to your providing your own preferred definition), ask the class what they think of the proffered definition and if they can improve upon it.

SCIENTIFIC ACCULTURATION PRINCIPLES

The first five experimental principles presented in this chapter are relatively self-explanatory and deal with (a) scientific acculturation (#1), (b) a cultivatable attribute (#2), and (c) the preparation necessary for conducting experiments (#s 3–5):

**PRINCIPLE #1:** Do not contemplate a scientific career if you are not prepared to be absolutely, uncompromisingly, unfashionably honest and transparent.

**PRINCIPLE #2:** Always be skeptical, but attempt to personally cultivate being open-minded.

**PRINCIPLE #3:** Do not contemplate independently conducting an experiment until you have mastered your general field and its experimental underpinnings.

**PRINCIPLE #4:** Do not contemplate conducting research whose primary purpose is to change the way clinicians practice their professions unless you are an experienced clinician yourself or have an active co-investigator who is.

**PRINCIPLE #5:** Conduct your first research forays under the tutelage of an experienced, principled mentor.

I think some class discussion on these points would be helpful although you may want to give your own take on them. Of the five, the first is repeatedly emphasized throughout the book as well as in the chapter on bias (Chapter Thirteen). I originally included others in this chapter (e.g., the necessity of working very hard) but made a decision to pare their numbers down as much as possible. I’m sure that you may have some favorites of your own to add; none, except perhaps the first one were inscribed on Moses’ stone tablets. I personally have numerous politically incorrect opinions (such as there are a lot more people conducting research than who are qualified to do so and that the majority of experiments conducted on humans are trivial in nature) but felt that it wasn’t my place to impose my biases upon your students.

SAMPLE ITEMS

*Stem for Q1 to Q6:* Which of the following studies would you classify as a prospective experiment? Why or why not? (These might be best delivered together before providing the answers, which
are of course keyed to the above definition. More items of this sort could be easily constructed from different disciplines by simply presenting the abstracts from a recent journal and having students identify which are experiments and which are not.)

Q1. Fourth-grade students are queried individually on their mistakes they made on a math problem administered to the class as a whole.

*Answer:* Not an experiment. No intervention was delivered.

Q2. Third- vs. fourth-grade mathematics classrooms are compared to ascertain the differences in the amount of reinforcement administered during a typical class period.

*Answer:* Not an experiment because no intervention was administered. (The fact that a comparison is made in a study does not imply that the study is experimental in nature.)

Q3. Psychology undergraduates are administered a personality inventory and then asked to fill out a scale assessing attitudes toward the propriety of specific actions taken to increase diversity in university settings to ascertain which (if any) personality factors were related to positive attitudes toward said actions.

*Answer:* This one is a bit more difficult because the personality inventory could be considered by some students as an intervention. However, there is no comparison that would permit the evaluation of the effects of this “intervention” on the outcome (attitudes).

Q4. Instead of being administered a personality inventory, half of the Q3 students were asked to read a passage making a positive case for educational diversity while the other half was given a passage pointing out the negative implications of institutional diversity actions in order to ascertain if the type of passage influenced attitudes toward diversity.

*Answer:* This one’s an experiment (although perhaps not a particularly meaningful one) because an intervention (the two types of passages) was implemented in order to measure the resulting effect on an outcome.

Q5. A teacher pretests her class prior to teaching a unit and then posttests it afterwards to ascertain the amount of learning she was able to elicit.

*Answer:* Yes, this one fits the criteria in our definition, although obviously it isn’t a particularly well-controlled experiment. The intervention was instruction in the unit (whatever that was) and the outcome was the difference between students’ pretest and posttest scores. This strategy will soon be graced with a name: the single-group pretest-posttest design, which is among the weakest of all experimental designs.

Q6. An educational researcher, wishing to learn whether charter schools produced more learning than regular public schools, accessed a state database that provided test score results for both charter and non-charter schools and statistically compared the two after controlling for previous student test scores.

*Answer:* No. Although two groups were statistically compared with respect to an outcome variable, no intervention was introduced prospectively.

For the next set of items, present the nine scientific functions that can be addressed experimentally (or have the class access them in their text). Like many of the other items in this manual, these are not multiple-choice or true-false, so you may want to present these in a class discussion format.
Stem for Q7 to Q15: The text claimed that eight of the following nine scientific functions were at least tangentially addressed by the Price et al. placebo study. Explain how each of these functions were and were not addressed (note that there are multiple correct examples of how some of these functions were met but I have simply indicated the ones I considered most salient). Also, and I promise not to keep saying this, feel free to substitute your own example.

Q7. Testing a theory (i.e., an experiment primarily concerned with testing the validity of a theoretical prediction by empirically illustrating the existence of a phenomenon and/or elucidating its etiology or the conditions under which it occurs).

Answer: The etiology of the placebo effect was demonstrated to be reinforcement of an expectancy effect (or conditioning). While there may be alternative explanations for the phenomenon, I can't think of any plausible ones given the number of alternative explanations eliminated by the study's careful laboratory conditions.

Q8. Determining the efficacy of an intervention (i.e., treatment, practice, strategy, procedure) by eliminating as many alternative explanations for this occurrence as possible.

Answer: A similar explanation to the first function if the manipulation of expectancy is conceptualized as an intervention (and there is no reason why it shouldn't be).

Q9. Improving the efficacy (e.g., by increasing the dose) of an intervention.

Answer: Three “doses” of expectancy were compared: strong, weak, and none (or negative), although perhaps this is somewhat of a stretch.

Q10. Comparing the efficacy of two (or more) interventions.

Answer: No, two different interventions weren't compared. An actual or hypothetical example from your discipline could probably be easily constructed or found in the literature. Alternately, the Stone, Grines, and Cox (2002) medical experiment in Chapter Seven ("Comparison of Angioplasty with Stenting, with or without Abciximab in Acute Myocardial Infarction") could be used (although the design may be a shade on the esoteric side at this stage in the course).

Q11. Replicating a previous finding.

Answer: A couple of instances occurred here, but the most notable was probably the fact that the memory of pain was more intense (following the manipulated expectation effect) than the participants actually experienced. This phenomenon had been demonstrated as a characteristic of a placebo effect in previous studies conducted by other investigators (e.g., the Amanzio, Pollo, Maggi, and Benedetti [2001] experiment, cited in Chapter Thirteen for any students who are interested in reading it [although this one was conducted following the Price et al. study]).

Q12. Determining how far a previous finding can be generalized (e.g., to other settings, populations, and conditions).

Answer: There have been quite a few demonstrations of the placebo effect both in laboratory and clinical settings. This study demonstrated its presence under the most controlled conditions up to its publication date of which I’m familiar. It might be interesting to ask students if they think this finding can be generalized to a clinical setting.
Q13. Translating an important observational finding into an experimental (therefore causal) one.

*Answer:* The placebo has been hypothesized and “observed” since Hippocrates’ time. This experiment constitutes a controlled (but far from the first) demonstration of the phenomenon.

Q14. Answering a new question generated by a previous finding.

*Answer:* There was some question as to whether the placebo effect for pain can be elicited solely through classic reinforcement in humans.

Q15. Challenging (or confirming) conventional knowledge (or practice).

*Answer:* This one confirms conventional scientific and public opinion that the placebo effect exists. A good deal of conventional knowledge has not, nor ever will be, confirmed (one example of which is that tiny needles can disrupt a physically undetectable energy source flowing through similarly undetectable meridians in order to relieve an incomprehensible medical diagnosis called gallbladder damp heat syndrome). It is difficult to demonstrate that something categorically does not exist, but not so difficult to experimentally demonstrate the possibility of an alternative explanation for their effects (e.g., that observed acupuncture effects may be more parsimoniously explained by the placebo effect).

**ADDITIONAL READING**

Should you wish to spend some additional time on these issues, the following readings might be assigned.


  This 63-page book (whose pages are small and the font large) comprises a highly recommended, in-depth treatment of scientific acculturation and ethics. It can be downloaded in pdf format at: https://download.nap.edu/login.php?record_id=12192&page=/download.php?record_id=12192. The book also contains an extensive bibliography and employs a number of brief case studies as a teaching tool that cover a variety of topics:

1. Advising and Mentoring
2. Mistakes and Negligence
3. Research Misconduct
4. Responding to Suspected Violations of Professional Standards
5. Authorship and Allocation of Credit
6. Conflicts of Interest

As such, it might serve as a good supplementary treatment of the topics covered in this chapter.
In this chapter I have attempted to frame the experimental process from the perspective of people’s everyday experiences in order (a) to illustrate that, as a species, we constantly engage in experimentation in our everyday lives (because making causal inferences is part of our everyday job description as humans) and (b) to introduce two inferential artifacts which people encounter in their decision-making process and which are, of course, applicable to the design of formal experimentation—as will soon become apparent.

The idea for introducing research design in this way came from a book I wrote for the public, entitled *Snake Oil Science* (also published by Oxford) which attempted to explain the experimental process to a lay audience. Some of the tables and figures in this and the next chapter also appeared in that book, which garnered a surprising number of positive comments from professionals regarding this informal approach to the logic of experimentation.

Referred to as impediments to making valid causal inferences in this chapter and promoted to the status of the first two formal experimental artifacts in Chapter Four, two important experimental concepts regardless of their designation are

**Natural history (or maturation),** which is defined as a tendency for some outcome variables to change over time irrespective of the introduction of an intervention or known extraneous events, and

**The placebo effect,** which is defined as a therapeutic response to an inert substance or irrelevant procedure occurring in individuals who believe they are receiving an effective treatment.

In my opinion it is necessary for students to understand the concept of experimental artifacts (or validity threats, threats to internal validity, or whatever terminology you prefer) and to be able to recognize them and their role in experimentation.
A number of seemingly built-in (or conditioned) psychological impediments were also discussed that conspire to make primitive experiments such as that performed by Mrs. Smith and Dr. Jones exceedingly poor foundations on which to base complex causal inferences—especially those in which the consequence does not immediately and inevitably follow the action—including:

1. Cognitive dissonance,
2. The personal importance we place on our beliefs,
3. The need to explain occurrences in the absence of evidence,
4. A tendency to ignore evidence that contradicts those beliefs,
5. A highly developed internal locus of control,
6. Respect for authority, and
7. A propensity to believe the absurd.

By implication these artifacts (and psychological impediments), to which many scientists are also vulnerable, require the design and conduct of more rigorous experiments than the everyday, informal one presented in this chapter. I have found that convincing people that their personal experiences often do not constitute a very reliable basis for making causal inferences is a tough sell, but it is an important concept to at least attempt to teach because this failing is one of the primary impetuses for conducting formal experimentation on humans. (A phenomenon related to this the unreliability of personal experience is the common unreliability of eye witness accounts reported in criminal trials.)

A key logical artifact that cannot be overemphasized (and that is especially relevant to many of the experimental artifacts to be introduced in the next chapter such as the co-occurrence of extraneous external events) is post hoc, ergo procter hoc: “after this, therefore because of this.”

### ACTIVITIES

Ask students to provide additional examples of some of the (a) logical (post hoc, ergo procter hoc), (b) physiological (natural history or maturation and the placebo effect), and (c) psychological (e.g., cognitive dissonance) impediments to making correct causal inferences discussed in the chapter. You may wish to prime the pump with some scenarios of your own and then ask students to identify the artifact, logical fallacy, or psychological impediment. Social and work-related interactions, past educational experiences with teachers, parental foibles, and the media could be all be used as sources.

### SAMPLE ITEMS

The next few questions are based on scenarios. I think they might be more effective if posed orally to the class and allow people to volunteer their answers.

**Stem for Q1 to Q4:** Suppose someone was suffering from heel pain and tried a sequential succession of treatments: medication from their internist, therapies prescribed by an orthopedic specialist, a podiatrist, a physical therapist, exercise, and a visit to an acupuncturist. Suppose that only following the acupuncture the pain completely disappeared. *(This experience was actually*
related to me by my internist when I once had the audacity to suggest to him that alternative medical therapies were primarily placebo effects. He immediately dismissed the contention, offering the experience of his own as proof that acupuncture was a great and true cure for pain.)

Q1. To what do you think the patient would ascribe the pain relief to?  
Answer: Acupuncture, since it was the last intervention introduced prior to the outcome change.

Q2. If this happened to you, to what would you ascribe the pain cessation?  
Answer: You have two options: The first, record the names of the students who answered acupuncture and subtract five points from their final grade. Or alternately, tell them to read Chapter Three and reread Chapter Two.

Q3. Based on the contents of Chapter Two, are there any alternative explanations for the cessation of the hypothetical individual's heel pain?  
Answer: Natural history, post hoc, ergo propter hoc, are acceptable, the placebo effect is not, because it wouldn't necessarily occur for one therapy and not the others unless the patient had a stronger belief in the final one tried than in the others. (There is some evidence that medical procedures [e.g., acupuncture, surgery] elicit stronger placebo effects than pills.)

Q4. Was the person seeking help for his or her heel pain really conducting an actual experiment as defined in Chapter One?  
Answer: Certainly he or she was conducting an experiment as we define it in the vernacular. It also meets one of the criteria mentioned in the definition (the prospective administration of an intervention—actually several of them). However, while outcome data were experienced by the investigator or patient, there is no indication that it was “collected” or “formally analyzed” anywhere other than in his or her fallible memory.

Q5. A mother teaches her preschool child how to read using materials supplied by an online company. During the course of this instruction, the mother felt that her child's speech (he initially stuttered slightly) improved somewhat. The mother ascribed the disappearance of the speech impediment to the company's materials. Are there any alternative explanations?  
Answer: Maturation or natural history is the most likely alternative explanation. So are (a) cognitive dissonance (e.g., the mother wanted to believe her financial investment paid off), (b) the need to come up with an explanation despite the lack of evidence, and, of course, (c) “after this, therefore because of this.” Other explanations might simply be that in the absence of objective measurement, the child's improvement may be a misperception on the mother's part. If anyone comes up with this one, compliment them and mention that this is one of the reasons we conduct scientific experiments in which change is measured objectively.

Q6. What could induce people to incorrectly think they received a benefit from a treatment provider?  
[A] Cognitive dissonance,  
[B] Respect for authority  
[C] The placebo effect  
[D] The tendency to ignore evidence that contradicts one's beliefs
Answer: [A] [B] and [D]. Not the placebo effect, because it involves an actual therapeutic benefit—not an imaginary one (i.e., the presence of the adverb incorrectly). However, if the item is interpreted as the patient actually received a benefit and that benefit was incorrectly attributed to the provider, then [C] is fine. Cognitive dissonance could operate if the individuals involved didn’t want to admit to themselves that they had wasted money needlessly. Respect for authority could operate here if the patient believed that his or her provider’s therapy was effective, as could ignoring evidence that contradicts our beliefs, especially since we know that symptoms are often worse in our memories than actually occurred (i.e., the Price study).

ADDITIONAL READING

Another discussion-related activity might be to assign an article by Chris Mooney (published by Mother Jones), entitled “The Science of Why We Don’t Believe Science: How Our Brains Fool Us on Climate, Creationism, and the Vaccine-Autism Link”: http://www.motherjones.com/politics/2011/03/denial-science-chris-mooney. This very readable article is most relevant to some of the psychological impediments discussed in the chapter and reinforces a point made several times in the text about the difficulties of convincing people (including scientists) on the basis of experimental evidence. The following link, to an article regarding Leon Festinger’s work, although a bit discouraging, should generate a lively discussion as well: http://www.motherjones.com/files/lfestinger.pdf. Both sources speak to the point that conducting a truly seminal experiment can have zero impact (at least initially) on changing people’s personal opinions.
CHAPTER 3

INTRODUCTION TO THE DESIGN OF MEANINGFUL EXPERIMENTS VIA THE CONTINUING ADVENTURES OF DR. JONES

This is a difficult chapter to teach, but one of crucial importance because it introduces (a) the design process via 10 basic decisions that must be made (although students can't be expected to understand the intricacies of most of them at this point), (b) the unavoidably subjective concept of meaningfulness (which is operationalized as much as possible via a series of behaviors for beginning investigators), and (c) the formulation of a research hypothesis (which can be conceptualized as a single-sentence synopsis (or operational definition) of the experiment itself.

WRITING HYPOTHESES

I fully realize that you may not employ hypotheses in your own research, perhaps preferring to formulate specific aims, purposes, or research questions. Pedagogically, however, I believe that formally stated hypotheses are important for beginners in both the design process and the evaluation of meaningfulness. I also think their formulation should be rigidly prescriptive with respect to their inclusion of the following five components, which can be used as a checklist when writing hypotheses:

- Intervention and its duration,
- Comparison of interest (e.g., single-group before-after or intervention vs. control contrast),
12 INTRODUCTION TO THE EXPERIMENTAL PROCESS

- Primary outcome,
- Type of participants, and
- Expected results.

SUGGESTED ACTIVITIES AND SAMPLE ITEMS

The topics contained in this chapter are so disparate that the sample items have been separated under these topic headings.

Practice writing hypotheses: I would suggest affording as much practice as necessary to the process of writing hypotheses. Abstracts from a journal in your area can be used for this purpose since they seldom contain formal hypotheses. These abstracts (or scenarios of your own making) can be assigned or performed in class (in groups or individually) and then shared with the class as a whole. A good starting point might be the research examples provided in this chapter and/or the Price placebo experiment presented in Chapter One. Alternately, scenarios can be provided:

Scenario for Q1 and Q2: A political scientist designed an experiment consisting of determining the effectiveness of two different written presentations favoring proposed legislation banning the ownership of pit bulls: one containing anecdotes of pit bull attacks on children and one involving statistics on the number and severity of reported attacks recorded in the state during the past 3 years. Undergraduate volunteers were randomly assigned to read one of the two arguments and then asked whether they would vote for the bill’s passage. (Scenarios such as this also can be used for class experiments involving students as subjects.)

One possible hypothesis for the study might therefore be, using the checklist just presented:

Undergraduate students [√ type of participant] randomly assigned to read a scenario involving graphic anecdotes [√ one of the interventions] favoring legislation to ban the ownership of dangerous animals will be significantly more likely [√ expected outcome] to report a propensity to vote for the legislation [√ primary outcome] compared to undergraduates randomly assigned to read a favorable scenario involving statistical information [√ the other intervention]. (√ The comparison of interest is implied since two randomized groups were specified.)

Q1. Can the reader pretty much glean what this experiment is about?
Answer: Hopefully yes, but improvements could be solicited from class.

Q2. Is this experiment meaningful?
Answer: I would answer “no” because of the outcome variable (reported voting behavior), but a political scientists might disagree, so there is probably no correct or incorrect answer here.

Scenario for Q3 to Q5: Use the video-gaming example in this chapter and write a hypothesis for it. One possibility might be the following:

Elementary school boys whose families have not purchased a video-game system but plan to do so [√ type of participants] and are randomly assigned to receive a gaming system as part of the experiment [√ the intervention] will achieve significantly lower scores [√ the expected result] on achievement tests [√ the outcome variable] than their randomly assigned counterparts whose
Q3. Does this hypothesis contain all five recommended constituents of a hypothesis?

*Answer*: Hopefully so, but again, improvements can be solicited from class. Admittedly hypotheses such as these can get a bit wordy, but their purpose isn’t to teach Hemingwayesque prose. It is to teach them how to design an experiment and ultimately assess its meaningfulness.

Q4. Is this experiment meaningful?

*Answer*: Yes, because it translates a potentially important correlational finding into an experimental one. (It is meaningful because school achievement in a societally and individually important outcome and anything capable of improving it is a potentially meaningful intervention.) Some students may object because the study is rather contrived and may not generalize (or they object that the results may not persist), but external validity (generalizability) is not part of this book’s meaningfulness criteria. Also, almost every experiment produces additional questions that need to be answered.

Q5. How could the study have been improved?

*Answer*: In my opinion it’s a pretty creative study, but a discussion here might provide some insights on the limits inherent in the experimental process. A follow-up interval would add some information on the duration of the negative effect but, given the delayed-intervention design, there wouldn’t be a useful comparison group available for evaluating the follow-up results.

**EVALUATING MEANINGFULNESS**

The first foray into the evaluation of experimental meaningfulness is introduced within the context of the design process itself, which primarily involves the first 3 of the 10 basic decision decisions:

1. The hypothesis (or hypotheses) to be tested,
2. The intervention to be employed (including the comparison group(s) to which it will be compared), and
3. The outcome that will be used to determine the intervention’s effect(s).

Specific suggestions relevant to these steps included (a) ascertaining whether the primary outcome is empirically related to the quality of human life and (b) the importance of (and approaches to) the development of viable interventions. The next evaluative steps include (a) the visualization of the most likely results to accrue from an experiment and (b) sharing the hypothesis (and the design that supports it) with other experimentally experienced individuals to obtain their opinions regarding scientific importance and procedural feasibility. These steps were considered important to deserve their own principles:

**PRINCIPLE #7:** Once the hypothesis is formulated, delineate all of the possible experimental outcomes, attempt to assess the probability of each of their occurrences, and attempt to visualize the actual numerical outcome likely to accrue if the hypothesis is supported.
PRINCIPLE #8: Always solicit as much professional feedback on your hypothesis, design, and rationale as possible prior to conducting an experiment.

Once the hypothesis and design have been evaluated from as many perspectives as possible, it is still important for students to understand that, in the end, the primary product of an experiment is “only” a “yes” vs. “no” or “supported” vs. “not supported” decision accompanied by a probability level. This may surprise some students, especially those who have read a number of experimental reports in which the discussion section encompasses a great deal more than the results of a simple hypothesis test. I believe that it is important to stress what really accrues from an experiment, which is basically a simple dichotomous answer to a straightforward question and not what this answer may lead to or the answer to myriad questions involving subsidiary or subgroup analyses (i.e., Principle #9, which by definition are not part of the experimental hypothesis).

Thus, once Principles #7, #8, and #9 have been adhered to, it is important to evaluate this relatively restricted product in light of Principle #10.

PRINCIPLE #10: Once a final hypothesis has been decided on, attempt to honestly assess your primary motivations (or aspirations) for conducting the experiment.

DISCUSSION AND ACTIVITIES SUGGESTIONS

As admitted earlier, this is a difficult chapter because the concept of scientific meaningfulness can be elusive and subjective. This in no way releases students about to design a dissertation (or investigators beginning a program of research) from considering the likely meaningfulness of these endeavors and whether or not the end product will meet their expectations and aspirations. It is the thoughtful, a priori consideration of what constitutes meaningfulness to the discipline and the investigator that is of the greatest potential utility of this evasive concept.

From a classroom perspective, about the best an instructor can do is to begin giving students practice in the process by assigning them the task of designing their own study.

1. One on a narrow topic that you provide the class as a whole, or
2. A follow-up experiment to one they have read or you have assigned.

SAMPLE ITEMS FOR DISCUSSION

Q6. What do you think about the author’s emphasis on scientific meaningfulness in general (e.g., do you think it is too restrictive or too nebulous)? (Please don’t e-mail me their answers!)

Answer: Obviously there is no correct or incorrect answer, so here are some follow-up questions.

Q7. Which meaningfulness component (e.g., outcome variables, interventions, visualization of results) do you think the author most explicitly defines and considers most important?

Answer: The outcome variable, although the stricture that it be causally related to a recognized societal or individual outcome requires a thorough command of the empirical literature.
Q8. Can you think of other acceptable, unacceptable, or excusable personal or professional motivations for conducting an experiment other than those posed?

*Answer:* Of course there are other acceptable (and pragmatic) motivations for conducting an experiment, such as testing a theory that one is skeptical about (or even conducting a preliminary test of one the investigator has developed). There are also occasions when curiosity simply gets the better of all of us and we have a real need “to see what would happen if . . .” Furthermore, most of us have to make a living, so there will be times (especially early in one's career) when a scientist has to take part in an experiment that he or she is convinced is useless. In such instances, I would argue, however, that one should (in as politically expedient manner as possible) at least (a) suggest alternatives and (b) object to violations of any of the 25 experimental principles presented throughout the book.

Q9. Do you think the author’s strictures against profit-related motives are too stringent?

*Answer:* First, there is nothing wrong with aspiring to be rich and/or famous. These are simply not very realistic scientific motives (although there are always a few Craig Venter–like exceptions). On the other side of the coin, a career in science may very well result in an economically comfortable lifestyle. The author’s main concern is that the first experimental principle (honesty and transparency) remain preeminent and never be violated for a profit motive. One reviewer of the book suggested that this particular principle be toned down (which it actually was), and obviously some frowned-on practices are worse than others. (Perhaps referring students to Chapter Thirteen, on bias, or using examples from your own experience and/or discipline would be helpful here. Also the mention of “seeding trials” in Chapter Thirteen might be illuminating. Alternately, you could simply tell students that you’ll return to these questions at the end of the course after they’ve read Chapter Thirteen.)

Q10. Is there anything wrong with (a) designing an experiment to serve as a marketing step for a product or (b) marketing an intervention that an investigator found to be effective?

*Answer:* Maybe yes, maybe no, but definitely maybe. And it is also definitely a path fraught with difficult decisions and troubling temptations. (As mentioned for the previous question, this issue can be revisited in Chapter Thirteen.) Designing an experiment in order to facilitate the marketing of a product can be suspect, but obtaining press coverage of experimental results is one way to assure the relevance of science to everyday life; many universities have public relations staff whose job it is to facilitate this. As a scientist, however, one must always avoid hyping findings beyond the limits of one’s data. (And in the presence of potential personal gain, this may be as difficult as a camel traversing the eye of a needle.)
CHAPTER 4

WHY POORLY DESIGNED EXPERIMENTS ARE INADEQUATE FOR MAKING COMPLEX INFERENCES

The Single-Group Pretest-Posttest Design (or the Unfortunate Conclusion of Dr. Jones’s Foray into the World of Science)

This is the final chapter in which design concepts are introduced within the context of Mrs. Smith and Dr. Jones’s life experiences. Hopefully, most students have found this a comfortable and low-key manner in which to be introduced to the experimental process.

This chapter is quite content dense regarding key experimental artifacts (which disparate disciplines describe as “threats to the internal validity of an experiment,” “experimental confounds,” or “confounding factors”). I personally prefer the term “experimental artifacts” to the more commonly used “threats to internal vs. external validity,” because the distinction between what is a threat to internal validity and one to external validity is sometimes blurred. (Many methodologists, for example, would classify artifacts #6 [volunteerism] and #9 [Hawthorne-like effects] as threats to external validity.) From my perspective, they are simply threats to the validity of the experimental inference and the distinction is not that important. If you disagree, your use of different categorizations and terms will not be contradictory to any of the other content in this or future chapters.

Note, also, that natural history and the placebo effect were introduced earlier as artifacts capable of invalidating the causal inferences we all make in our everyday life, but here it is important to reintroduce them in terms of artifacts or validity threats that have the potential of (a) being mistaken for, (b) inflating, (c) deflating, or (d) disguising a true experimental effect.
INTERNALIZING THE 10 EXPERIMENTAL ARTIFACTS AND THEIR AMELIORATION

Students should eventually commit these artifacts and their definitions to memory and it is to facilitate movement toward this end that the "sample items" are directed. Hopefully, they will also become internalized as different designs and control strategies are introduced in the course of the next several chapters. The 10 experimental artifacts introduced to this point, along with their definitions, include the following:

**#1**: *Natural history or maturation*: a tendency for participants' outcome values to change over the experimental interval as a function of their physical or mental conditions, time, or development.

**#2**: *Placebo effect*: a therapeutic response to an inert or irrelevant substance or procedure occurring in participants who believe they are receiving an effective treatment.

**#3**: *Regression to the mean*: a statistical tendency for participants selected on the basis of extreme values on the outcome variable to exhibit less extreme scores when remeasured.

**#4**: *Repeated testing*: a tendency for participants to exhibit higher (or more socially acceptable) scores the second time they are administered certain types of outcome measures, resulting from familiarity with the measures themselves (e.g., by correcting their initial mistakes or becoming aware of the intent of the questionnaire).

**#5**: *Instrumentation changes during the course of an experiment*: changes on an outcome variable attributable to changes in the format of the measure itself (or the manner in which it is administered or the data are collected).

**#6**: *Volunteerism*: a tendency for the participants (who are almost always volunteers) to take part in an experiment to possess a greater or lesser propensity to change on the outcome variable than those who declined to participate.

**#7**: *Extraneous external events (aka history)*: the potential for an event external to the experimental setting to occur during the course of the experiment capable of independently influencing responses on the outcome variable.

**#8**: *Demand characteristics (or the "good-subject" phenomenon)*: a tendency for participants to provide the investigator with desirable (or socially acceptable) outcomes following the intervention.

**#9**: *The Hawthorne (or reactivity) effect*: a tendency for participants to behave differently as a function of being in an experiment or being observed.

**#10**: *Experimental attrition*: the possibility that participants who withdraw from experiments may (or would) have changed more or less on the outcome variable than those who remain.

**SUGGESTED ACTIVITY**

Of the 10, I have found the most intuitively difficult artifact to truly understand is regression to the mean. The following activity is designed to help in this regard. (The idea for this activity was adapted from Palmerino, C.C., King, J.C., Gordis, F.W., & Cozby, P.C. [1997]. *Instructor's Manual to Accompany Methods in Behavioral Research* [6th ed.]. Mountain View, CA: Mayfield.)
Experiments Are Inadequate for Making Complex Inferences

1. Secure 100 small index cards from the supply cabinet (or 50 if times are tight and cut them in two). Write the numbers 1 to 100 on each card and shuffle them well (preferably in front of the class).

2. Next deal each member of the class the same number of cards. For example, if you have 20 students, deal them 5 cards each. If you have 23, deal them 4 cards each. If you have less than a dozen students, this activity won’t work unless you deal each student two or more hands.

3. The object of this game, you can tell them, is to be dealt the highest cards defined as the total sum dealt.

4. Once everyone has summed their cards, write their scores on the board, from highest to lowest. Write the names of the students with the highest four or five scores and the lowest four or five scores next to their numbers. (If the class is small, compensate by selecting two or three scores to represent each tail of the distribution.)

5. Add up each group of five scores and write them on the board.

6. Shuffle and deal the cards again and repeat the scoring process.

7. Ignoring everyone’s scores, except those of the individuals whose names were written on the board, corresponding to the highest and lowest sums.

8. Write these individuals’ second sums next to their first sums, add them, and place the four sums in a table such as the following:

<table>
<thead>
<tr>
<th></th>
<th>First Trial Scores (or Pretest)</th>
<th>Second Trial scores (or Posttest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest scoring individuals first Trial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highest scoring individuals first Trial</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Regression to the mean should occur with the initially lowest “scoring” students’ improving on the second trial and the reverse occurring for the other group. I think what is useful about this demonstration is the illustration of the universality of the regression to the mean phenomenon. Some discussion of other possible scenarios might prove interesting, such as: When two extremely tall parents have a child, does this child, on average, wind up being as tall as their parents (taking gender into account and assuming the children had the same nutritional advantages as their parents)? (Note that the child is still likely to be taller than the population average, just not as tall as his or her parents.)

**SAMPLE ITEMS**

The following questions should be answered by selecting either (a) the most applicable artifact from the 10 possibilities presented below or (b) all that apply (as indicated in the question stems). Some of the provided answers are challengeable, since there are often situational exceptions to the circumstances under which these artifacts occur, but the challenge process itself is a good way for students to internalize the concepts themselves. (*They should assume a single-group design has been employed.*)

#1: Natural history or maturation
#2: Placebo effect
#3: Regression to the mean
#4: Repeated testing
#5: Instrumentation changes during the course of an experiment
#6: Volunteerism
#7: Extraneous external events (aka history)
#8: Demand characteristics (or the “good-subject” phenomenon)
#9: The Hawthorne (or reactivity) effect
#10: Experimental attrition

Q1. Which artifact is the easiest one to avoid?

*Answer:* Instrumentation changes (#5), because there is seldom any need to change the manner in which the outcome is measured.

Q2. Which artifact is the most difficult to avoid?

*Answer:* Volunteerism (#6), because—with a very few specialized exceptions—individuals must volunteer for an experiment and sign an informed consent statement.

Q3. Which five artifacts are more likely to produce false positive than false negative results?

*Answer:* Placebo effect (#2), regression to the mean (#3), repeated testing (#4), demand characteristics (#8), and Hawthorne/reactivity-type effects (#9), because all five are more likely to produce effects in the hypothesized or socially desirable direction unless the outcome variable itself reflected a societally or personally negative characteristic or behavior. Regression to the mean, because in most cases participants are selected because of their extreme scores on the outcome variable and the experimental purpose involves moving said participants’ scores toward “normalcy.”

Q4. Assuming conditions favor its occurrence (and it occurs with some frequency), which artifact is potentially most damaging in a single-group design?

*Answer:* Opinions may differ here, but, if it occurs, I would make a case for attrition, simply because of the plausibility that people who leave an experiment are in some way different from those who do not.

Q5. Which artifact is least likely to occur if the participants are selected on the basis of their representativeness (or lack thereof) to the population as a whole?

*Answer:* Regression to the mean, because it only occurs when participants are selected (or select themselves) on the basis of having extreme scores on the outcome variable.

Q6. For which three artifacts does the probability of occurrence increase most dramatically as the experimental interval itself increases?

*Answer:* Natural history or maturation (#1) because there is more time for changes in the outcome variable (or in the participants themselves), extraneous external events (#7), because there is more time for such untoward events to occur, and experimental attrition (#10) for the same reasons and because of either greater demands on the participants or participant relocation for noncaptive audiences.

Q7. Which artifact, if any, do you find most counterintuitive (“none” is an acceptable response)?

*Answer:* While there is, of course, no right or wrong answer here, most students will select regression to the mean. Asking those who do select this option *why* might be a good teaching point.
Experiments Are Inadequate for Making Complex Inferences

This diagnostic question should be followed up via remediation, perhaps by forming groups of students who select the same artifact and having students who selected "none" work with them.

Q8. Two additional experimental principles were introduced: the 11th (avoiding interventions whose relationship to self-reported, or reactive observational, outcome variables are potentially obvious to participants) and the 12th (to the extent permitted ethically and scientifically, communicating as little as possible to participants and research staff about the specific purpose and hypothesized results of an experiment). To which four artifacts do you suspect this advice is most relevant?

Answer: Demand characteristics (#8) and Hawthorne/reactivity-type effects (#9) are the most obvious ones. The intervention itself may induce a placebo effect (#2) by imparting expectations of benefit. Perhaps less likely, a repeated testing (#4) could be exacerbated by the contraindicated conditions.

Q9. Suppose an instructor administered a brief quiz on experimental artifacts at the beginning of the class and then administered the identical quiz at the end, following a lecture on the topic, in order to test the effectiveness of his or her instruction. Assuming students scored significantly higher on the posttest than on the pretest, explain why each of the 10 artifacts would be likely or unlikely to invalidate the inference that the instruction was effective.

Answer:

#1: Not natural history or maturation, because the interval was too brief.
#2: Not the placebo effect, because there is no placebo effect for learning.
#3: Not regression to the mean, because the class membership wasn't selected because of their unusual ignorance of experimental design concepts (although you may suspect it was somehow).
#4: Repeated testing is a definite possibility, since students who didn't know the answer to a particular question would recognize what their errors were and therefore be in a position not to make the same mistakes twice.
#5: Not instrumentation changes, because the same quiz was administered under the same conditions at both pretest and posttest. However, if the instructor told students that the posttest would count on their grades as she or he passed it out, then this instrumentation change might definitely affect performance.
#6: Not volunteerism, because the quizzes were required.
#7: Not extraneous events (aka history), because it is hard to imagine how an event could occur that would increase student attention or learning unless some distraction occurred during the pretest. (The instruction delivered is the intervention, not an extraneous event.) If the results had been negative, this might have been a more plausible possibility (e.g., a jackhammer being used outside the building during instruction, a malfunctioning thermostat, and so forth).
#8: Not demand characteristic, because supplying the correct answers to test items can't be "faked."
#9: Not Hawthorne or reactivity effects, because the participants themselves weren't being observed, and even if they were, it wouldn't affect learning unless it induced students to pay more attention to the lecture and/or reduce disruptive behaviors.
#10: Not experimental attrition, unless some students left the class in protest (which is a bad omen for this particular class) because they didn't like quizzes.
Q10. Suppose a single-group experiment involved a therapist who wished to assess the effectiveness of a new component added to his or her treatment of patients with depression. For new patients, a brief assessment of the degree of depressive symptoms is administered during the first session, the new component is introduced, and the patients are reassessed at a later session. Assuming positive results were obtained, list an alternative explanation for these results that each of the 10 experimental artifacts might conceivably engender.

Answer:

(#1) **Natural history** is a possible alternative explanation (since depression waxes and wanes naturally and patients might have sought relief when symptoms were at their worst).

(#2) **The placebo effect** could be operative (because hopefully the therapist did all he or she could to plant the expectation of relief).

(#3) **Regression to the mean** could also be problematic, since only the more depressed patients (or those for whom depression was greater than normal for them) would be the ones most likely seeking treatment.

(#4) **Repeated testing** is possible but less likely here because the patients may see it in their best interest to respond as accurately as possible to the assessment. The artifact might still operate weakly for those patients who realized they did not understand the questions the first time around (or who were sensitized to the meaning of the terms implied by the therapy itself).

(#5) **Instrumentation changes** most likely would not occur if a written questionnaire was used and administered by the same individual. However, if the therapist administered the assessment orally, subtle changes in affect due to his or her increased familiarity with the patient (and unconscious communication of his or her hope for improvement) might induce positive responses. (Or, of course, if a questionnaire was used as the pretest and oral administration was used later, then instrumentation changes would be a definite threat).

(#6) **Not volunteerism**, since all the therapist's patients "volunteered" for help and the therapist did not aspire to generalize the results past his or her patient population.

(#7) **Extraneous external events** might occur over time if, for example, most of the original assessments were conducted in winter and the reassessments were conducted during the spring (since there is a seasonal component to depression).

(#8) **The good-subject phenomenon** might induce some patients to report fewer symptoms following treatment as this is obviously what the therapists hopes will occur and the patient may desire approval or fear rejection.

(#9) The Hawthorne effect probability isn't applicable here (although it is possible to visualize how reactivity could influence the posttest as a function of the therapist's continual insistence on compliance with the therapeutic regiment or continually asking the patients if they have improved on this or that dimension).

(#10) **Attrition** could influence the results if a number of patients drop out of therapy prior to the second assessment and if this were because they felt they weren't being helped.
Arguably the two most important experimental strategies in human experimentation involve

1. The use of a control/comparison group and
2. The random assignment of participants to that control/comparison and an intervention group.

I don’t think these two connected strategies (as well as their underlying concepts) can be overemphasized in experimental design. This is especially true of random assignment, which approaches the status of a religious tenet for research methodologists and successful investigators.

Of the two concepts (the use of comparison groups and randomization of participants thereto), the logic behind the use of a control group tends to be intuitive and already internalized, while the concept of randomization (which involves the almost cynical acceptance of and subsequent capitalization on ignorance) is counterintuitive and not well understood by many.

(Random assignment was defined in terms of ensuring that each participant possesses an equal chance of being assigned to each experimental group compared to all other participants.)

**RANDOM ASSIGNMENT ACTIVITY**

*Step 1:* Have your class count off consecutively from front to back with each person writing his or her number down on a scrap of paper.

*Step 2:* Have each student write down a number from 0 and 10 on their slip of paper containing their ID. If the class is small, have the students write down a second number from 11 to 20. (Explain that this second step is necessary because the utility of random assignment increases as the N/group increases).

This number can represent anything you wish, to give the data some context. I’d suggest that you identify it as a pretest value representing the hypothetical study’s outcome variable, which I’ll simply call “propensity to change on the outcome.”
Step 3: Access the Internet and choose a random assignment calculator. (Or use your statistical package if you prefer.) One site is http://www.graphpad.com/quickcalcs/randomize1.cfm

You should see the following entry line:

Assign ___ subjects to each of ___ groups. Repeat 1 time.

Fill in the desired N/group (e.g., if there are 24 students in the class the N/group will be 12) in the subject box and 2 in the group box. Your output will be in terms of A or B, so assign one as the experimental group and one as the control if you wish. If you prefer, flip a coin; it doesn’t matter as long as you assign the group identification before you view the results—but stress that you’d never do this for an actual experiment. Let’s pretend that the following results were obtained:

<table>
<thead>
<tr>
<th>ID#</th>
<th>Group Assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
</tr>
<tr>
<td>13</td>
<td>B</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
</tr>
<tr>
<td>18</td>
<td>B</td>
</tr>
<tr>
<td>19</td>
<td>B</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
</tr>
<tr>
<td>21</td>
<td>A</td>
</tr>
<tr>
<td>22</td>
<td>A</td>
</tr>
<tr>
<td>23</td>
<td>B</td>
</tr>
<tr>
<td>24</td>
<td>B</td>
</tr>
</tbody>
</table>
Enter the Control Group

(If the class has an uneven enrollment, say, 25, type in 13 in the subject box and simply ignore ID #26 since it doesn’t exist.)

Step 4: Access (or have a student access) a t-test calculator, such as http://graphpad.com/quickcalcs/ttest1.cfm. (If possible, project image for the class to review.)

Step 5: Have the A group (or however it is designated) read off the numbers they have written on their sheet of paper while the data entry student enters the data into the calculator. Repeat the process for the second group. Let’s pretend that the following numbers and results accrued.

A RANDOM ASSIGNMENT EXAMPLE USING AN ONLINE CALCULATOR

Choose data entry format

| Enter up to 50 rows.        | Unpaired t-test. |
| Enter or paste up to 2000 rows. | Welch’s unpaired t-test (used rarely). |
| Enter mean, SEM and N.      | Paired t-test.   |

Choose a test

Enter data

Label A B

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

The results of interest for present purposes are asterisked and bolded:

Unpaired t-Test Results

p-value and statistical significance:

*The two-tailed p-value equals 0.9459.
Add-on activity: (This one may be a bit statistical in nature, but it is intended to cement the concept of random assignment.) A potentially entertaining out-of-class activity is to promise students some reward (extra credit or your undying esteem) if they can legitimately subvert the random assignment process in no more than two attempts. Use an even number such as 40 and allow them to employ a number range of 20. Remind them, however, that the values shouldn’t be dichotomous and should be more or less normally distributed prior to randomization. If this is getting ahead of their statistics, simply tell them they should use at least half of the 20 possible values. (Remind them that once they’ve chosen their numbers they must randomly assign them to two groups through the use of the online calculator and then assess whether or not they succeeded in beating the randomization process via the \( t \)-test calculator.) Some may succeed in obtaining statistical significance, so allow them to share their strategy with the class. You can then make the determination of whether or not their results were due to chance. If you don’t know, replicate their strategy with another set of numbers. (Tell them that later independent replication of a finding will be identified as the best way to ascertain if experimental results are due to chance or bias of some sort, although of course this isn’t an experiment as defined previously.) Discuss reasons why some members of the class obtained statistical significance and some did not. (Five percent of the class’s overall attempts should have obtained statistical significance.)

**Types of control/comparison groups**

Five types of control/comparison groups were discussed: (1) no treatment, (2) placebo, (3) attention, (4) treatment as usual/standard care, and (5) alternative treatments. Students should understand the difference between all five along with their situational strengths and weaknesses. It is difficult to overemphasize the importance of selecting and formulating an appropriate control/comparison group in terms of ensuring (a) its fit with the experimental purpose and intervention, (b) the avoidance and control of as many experimental artifacts as
possible (in order to arrive at a final, valid inference regarding the efficacy and non-efficacy of the key experimental conditions), and (c) the experiment’s ultimate meaningfulness.

**SUGGESTED ACTIVITIES**

*Suggestion #1:* I think everyone agrees that the best way to learn research is to actually conduct it, but since that’s not feasible for classroom instruction, the next best way is to read and evaluate as many published experiments as possible. So if you haven’t already begun, from this time on I’d suggest that you have students read experiments of your choice and theirs and respond to questions (or write critiques).

For the present topic (types of control groups), assign students (or groups thereof) to find an example of each of the five genres in their discipline (or a related discipline in which they are interested). Have them abstract the study very briefly (or cut and paste the published abstract) and evaluate the appropriateness of the control group and the authors’ interpretation of the results. (For some disciplines a pure placebo group may be quite difficult to find, so if this is true of yours adjust the assignment accordingly.)

*Suggestion #2:* If *Suggestion #1* is not feasible, have students construct hypothetical experiments using each of four of the five comparison groups (since placebo and attention controls are seldom both applicable for the same experimental scenario) and then rate them based on the 10 artifacts listed earlier. Once completed, have them rate the artifacts that might have been operative if they had simply employed a single-group pretest-posttest design.

<table>
<thead>
<tr>
<th></th>
<th>Single Group Pretest-Posttest</th>
<th>No Treatment Control</th>
<th>Placebo Control</th>
<th>Attention Control</th>
<th>Standard Care</th>
<th>Alternative Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selection</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Natural History</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Placebo Effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Repeated Testing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Extraneous Events</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Demand Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reactivity Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Attrition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Differential Attrition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Believe it or not, I labored long and hard trying to decide whether to include these charts and ratings in this and the next chapter. So many situational exceptions exist that few of them are absolute and unchallengeable under special circumstances. Placebo controls, for example, are given the only unqualified “√” for controlling the placebo effect, but that generalization is based on the assumption that the placebo procedure is properly constructed—which often isn’t the case even in drug trials. All randomized control groups are likewise given a “√” with respect to the control of extraneous external events, but unusual, unanticipated circumstances could arise for an individual experiment in which such an event occurred in one group and not another. Many of these possible exceptions can be anticipated or detected by carefully monitoring the conduct of a trial, but that is the subject matter of Chapter Ten. (Also, some artifacts can be prevented irrespective of the design employed, by use of the ameliorative strategies suggested throughout the text.)

Finally, as noted in the text, some of the control/comparison group labels employed here are (a) not mutually exclusive, (b) subject to subjective interpretations, and (c) given different names in different disciplines. The same comparison group, for example, might be labeled as an attention control, alternative-treatment, or a standard-care group by different investigators or research consumers depending on their orientation, thus these ratings must be interpreted cautiously. I therefore apologize but you will have to decide for yourself whether they are worth the effort or actually serve their primary objective: to encourage students to “think” in terms of evaluating design decisions in terms of the inferential artifacts that bedevil experimental research.

Since the virulence of these artifacts is experimentally situational, let’s assume that the following experiment was posed, which was constant except for the control/comparison group employed.

Elderly assisted living residents are randomly assigned to attend a 30-minute health education class once a week for 6 weeks. The outcome variable involves pre-posttest changes on the adoption of the preventive behaviors taught. The control/comparison groups were (a) a no-treatment control in which the facility agreed to suspend all programs and related activities during the 6-week experimental period, (b) a 30-minute class on non-health topics (attention control), (c) the usual health promotion suggestions provided by the facility staff as well as a menu of other activities and courses provided by the facility (standard care), and (d) once-a-week seated yoga exercises (alternative treatment). Note that these ratings, as always, are influenced by opinions regarding the situation, hence some artifacts are rated as slightly more problematic because of the close proximity of control and intervention participants, while the alternative treatment (because of its potential attractiveness) is rated higher on some.

<table>
<thead>
<tr>
<th>Table: Control Groups</th>
<th>(a) No-Treatment Control</th>
<th>(b) Attention Control</th>
<th>(c) Standard Care</th>
<th>(d) Alternative Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Natural History</td>
<td>√</td>
<td>√</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Placebo Effect</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Regression to the Mean</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Repeated Testing</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
Note that all four of the control/comparison groups are effective in preventing selection and controlling for regression to the mean, repeated testing, and extraneous co-occurring events. Note also that the latter—if sufficiently disruptive—is still capable of producing either a false positive or negative result. The other artifacts, when applicable, are best controlled by the procedural suggestions tendered in the text.

### SAMPLE ITEMS

**Q1.** Does random assignment ensure that equal numbers of participants will be assigned to groups? [Y/N]

*Answer:* No, unless the investigator constrains the assignment in some way. The random-assignment calculator used earlier forces the groups to have equal numbers but pure random assignment does not. (It is, however, preferable to force the groups to have equal numbers of participants.)

**Q2.** Does random assignment guarantee that the experimental groups do not differ in some significant way? [Y/N]

*Answer:* No, it only makes initial nonequivalency improbable.

**Q3.** Can random assignment be used for more than two groups? [Y/N]

*Answer:* Yes, as many groups as are being investigated. The same logic and definition apply.

**Q4.** In terms of the 12 artifacts listed to this point, which is the only one that the random assignment of participants to two or more groups helps prevent (as opposed to help control) irrespective of the type of control/comparison group employed?

- [ ] Natural history or maturation
- [ ] Placebo effect
- [ ] Regression to the mean
- [ ] Repeated testing
- [ ] Instrumentation
- [ ] Volunteerism
- [ ] Extraneous external events (history)
- [ ] Demand characteristics (or the “good-subject” phenomenon)
- [ ] The Hawthorne (or reactivity) effect
- [ ] Attrition
- [ ] Selection
- [ ] Differential selection

---

<table>
<thead>
<tr>
<th>Extraneous Events</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demand Characteristics</td>
<td>X</td>
<td>≈</td>
<td>≈/X</td>
<td>Y</td>
</tr>
<tr>
<td>Reactivity Effects</td>
<td>X</td>
<td>≈</td>
<td>≈/X</td>
<td>Y</td>
</tr>
<tr>
<td>Attrition</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Differential Attrition</td>
<td>X</td>
<td>√/≈</td>
<td>√/≈</td>
<td>Y</td>
</tr>
</tbody>
</table>
InTroducTIon T o THE EXPErIMEnT AL ProcEss

Answer: [#11] Selection, which was defined as the possibility that the experimental groups initially contained participants with different propensities to change on the outcome (or to be differentially responsive to the intervention[s]).

Q5. Assuming the use of an appropriate control/comparison group, which of the remaining 11 experimental artifacts does the random assignment of participants facilitate procedurally or statistically controlling?

[#1] Natural history or maturation
[#2] Placebo effect
[#3] Regression to the mean
[#4] Repeated testing
[#5] Instrumentation
[#6] Volunteerism
[#7] Extraneous external events (history)
[#8] Demand characteristics (or the “good-subject” phenomenon)
[#9] The Hawthorne (or reactivity) effect
[#10] Attrition
[#11] Differential selection

Answer: [#s 1, 2, 3, 4, 7, 8, 9, 11] Assuming the use of an appropriate control/comparison group, random assignment helps prevent only selection, but it does facilitate the procedural control of the effects of 8 of the other 11 artifacts, assuming the use of an appropriate control/comparison group. Natural history, the placebo effect, regression to the mean, repeated testing, the co-occurrence of extraneous, demand characteristics, reactivity effects, and differential attrition can theoretically be procedurally controlled (or subtracting out) in the sense that they theoretically occur equally in each group when said groups are appropriately constituted. (Note the huge qualifier here: assuming the use of an appropriate control/comparison group.) The placebo effect, demand characteristics, reactivity effects, and differential attrition require a specially constructed control/comparison group when they are operative. Nondifferential attrition is not normally associated with the type of control group but dependent on the amount of effort required by the experiment, its duration, and other characteristics (e.g., affective, demographic) associated with the participants and experimental context. Instrumentation changes must be prevented procedurally and volunteerism is not completely preventable in prospective research involving the need for informed consent, although it can be reduced via use of a seldom-used design introduced in the next chapter.

Q6. Which of the following artifacts is a no-treatment control unlikely to be able to prevent or control in experiments vulnerable to them?

[A] Selection
[B] Natural history
[C] Differential attrition
[D] Extraneous events
[E] Demand characteristics

Answer: [C] and [E] Randomization of any type of comparison group will largely prevent selection and control natural history and co-occurring extraneous events. If the experimental conditions are vulnerable to differential attrition (e.g., in cases where the study interval is long, the intervention is quite attractive, or a good deal is required of participants), then its potential
is exacerbated by an unattractive control group. The same is true for demand characteristics (e.g., investigators or study staff who communicate their aspirations to participants) and reactivity effects (e.g., when interventions are employed which quite obviously match self-reported outcome variables).

Q7. Which two types of comparison group are the most difficult to constitute?
   [A] No-treatment control
   [B] Placebo control
   [C] Attention control
   [D] Standard treatment
   [E] Alternative treatment

Answer: [B] and [C] Credible placebo and attention placebo controls are “incredibly” difficult to construct and beginning investigators would be wise to employ only those that have been validated previously in a published study. (And even then they would be wise to enlist the help of a co-investigator who was involved in its construction and implementation.) A standard-care comparison group can be quite challenging to operationally define as well if it is not already in place, thus [D] is not a bad answer.

Q8. In which type(s) of comparison group(s) is differential attrition most problematic?
   [A] No-treatment control
   [B] Placebo control
   [C] Attention control
   [D] Standard treatment
   [E] Alternative treatment

Answer: No-treatment control followed (in order) by standard treatment and alternative treatment. If the attention control is especially transparent, it may produce more attrition than either standard-care or alternative-treatment conditions.

Q9. What characterizes a single-blinded design?
   [A] A placebo or attention control is required.
   [B] A single-blinded design disguises the group assignment from either the participants or the experimental staff interacting with the participants.
   [C] A single-blinded design disguises the group assignment only from the participants, not from the experimental staff interacting with the participants.

Answer: [C] Single blindness normally refers to blinding the participants. Placebos or attention placebos are often used in this type of study, but alternative-treatment controls can also be used if the informed consent is carefully worded.

Q10. In a placebo-controlled trial with a well-constructed and implemented placebo, which typically exhibits a placebo effect?
   [A] Intervention group
   [B] Placebo group
   [C] Neither
   [D] Both

Answer: [D] Both—hopefully equally, but some participants usually correctly guess their group assignment. Even when the intervention is effective, it will still have a placebo component.
NEWLY INTRODUCED RANDOMIZED DESIGNS

Writing a textbook on a subject is different from (and easier than) teaching it. However, one constraint that is difficult to overcome in writing prose is how to stress certain points without compromising the flow of the narrative. This and the preceding chapter deal with content that is among the most important in the book, and I therefore feel the need to stress the importance of some of the points that I think should be emphasized in class, even if it means devoting extra time to the content in question (perhaps even to the detriment of covering some of the later subject matter).

In effect, this chapter is an extension of Chapter Five, which introduced the key concepts of control groups, random assignment, and the more common experimental artifacts that need to be carefully considered during the design phase of an experiment. Chapter Six continues this discussion by introducing different types of experimental architecture in common use in most disciplines. These specific designs (along with the types of control/comparison group chosen) impact the inferential validity of an experiment. To facilitate this discussion, certain contrasting nonrandomized designs (some commonly used, some not) are discussed to solidify the importance of random assignment and the key role played by the choice of a control/comparison group.

For your convenience, I have reproduced the actual design figures in case you wish to copy them for presentation on the board:

1. The randomized pretest-posttest design, which basically combines a single-group pretest-intervention-posttest with a pretest-control/comparison group-posttest architecture (see Figure IM.1). (Type of control/comparison groups chosen is optional and situational.)
InTroducTIon T o THE EXPErIMEnT AL ProcEss

4. Randomized delayed-intervention designs (Figure IM.4), in which one condition receives the intervention (immediately after randomization) while the other group receives the intervention following end of treatment (EOT). This design can involve a pretest, covariate, or both, as well as any type of control/comparison group deemed appropriate.

5. The randomized cluster design, which simply means that clusters of individuals (e.g., classrooms of students, schools, treatment facilities) are randomly assigned to
The three quasi-experimental designs (the nonrandomized pretest-posttest, the separate-cohort design, and patient preference designs) were introduced primarily to serve as a contrast to the randomized designs just depicted. Of the three, the first is the most commonly employed and the only one that most students will encounter or use in their own work, although the separate-cohort strategy has its own charms. Architecturally, of course, the nonrandomized pretest-posttest group design (Figure IM.5) is almost identical to its randomized counterpart.

The differences between the randomized and nonrandomized designs should be stressed. It is counterintuitive to some students that the primary advantage of the former over the latter resides in the ability of randomization to prevent selection. With that said, the appropriateness of a design can be dependent on the specific intervention, outcome variable, sample size, control/comparison group characteristics, and research hypothesis employed.

![Randomized Delayed-Intervention Design](image1)

![Nonrandomized Pretest-Posttest Control Group Design](image2)

**QUASI EXPERIMENTAL DESIGNS**

The three quasi-experimental designs (the nonrandomized pretest-posttest, the separate-cohort design, and patient preference designs) were introduced primarily to serve as a contrast to the randomized designs just depicted.

Of the three, the first is the most commonly employed and the only one that most students will encounter or use in their own work, although the separate-cohort strategy has its own charms. Architecturally, of course, the nonrandomized pretest-posttest group design (Figure IM.5) is almost identical to its randomized counterpart.

The differences between the randomized and nonrandomized designs should be stressed. It is counterintuitive to some students that the primary advantage of the former over the latter resides in the ability of randomization to prevent selection. With that said, the appropriateness of a design can be dependent on the specific intervention, outcome variable, sample size, control/comparison group characteristics, and research hypothesis employed.

**FOUR EXPERIMENTAL ARTIFACTS**

By itself, randomization of participants to groups does little to prevent the four artifacts (the first two of which are the more serious and therefore deserving of more attention) presented in this chapter:

*#13: Other experimental confounds:* It is worth repeating that the prefixed adjective “other” acknowledges the fact that artifacts such as selection, natural history, and the placebo
effect can all be considered experimental confounds. These particular confounds have been given their own names here solely for pedagogical and mnemonic purposes.

**#14: Treatment diffusion:** This artifact is best controlled via a randomized cluster design, although the nonrandomized pretest-posttest design employing intact groups (with all of its attendant disadvantages) can avoid participants in the different experimental groups being in close enough proximity to communicate with one another.

**#15: Resentful demoralization:** This artifact’s greatest actual threat may reside in its potential for increased attrition among control participants (i.e., differential attrition), although it could conceivably negatively affect self-reported outcomes. Of the randomized designs presented, the delayed-intervention and (possibly) the randomized cluster designs provide the greatest protection against its effects, although some outcome variables and participant samples are undoubtedly immune to this artifact. In my opinion (i.e., I have no evidence for the statement), resentful demoralization is not a particularly virulent threat in most experiments.

**#16: Compensatory rivalry:** There is little, if any, controlled evidence that this artifact actually exists with any frequency, but the possibility of its occurrence can be prevented by blinding patients and service providers. Its likelihood, low to begin with, can also be reduced by using a delayed-intervention design.

### A CRUCIAL EXPERIMENTAL PRINCIPLE

The following design principle, buttressed by #13 (the importance of randomly assigning participants to one or more control/comparison groups), is among the most important proffered in the text, so if you feel I have failed to give it the emphasis it deserves please compensate for me.

**PRINCIPLE #17:** Prior to conducting an experiment, always (a) identify all potential confounds, (b) eliminate those which are plausibly related to the study outcome by standardizing and/or counterbalancing procedures across experimental conditions to the maximum extent possible, and (c) monitor the effects of those confounds that cannot be eliminated or counterbalanced.

### SUGGESTED ACTIVITIES

Assign groups of students to locate a published example of one of the designs presented in this chapter that is not uncommonly employed in your discipline. (You might want to suggest some specific journals for them to search.) Have them critique the strengths and weaknesses of the design and its implementation with special reference to the 16 artifacts listed in the chart that follows. This chart could be accompanied by a dichotomous scoring system (\(\times\) = problematic in your opinion; \(\square\) = not problematic). Feel free to delete any artifacts you think are not endemic to your discipline. Naturally, there will be some subjectivity in the assignment of ratings here, so tell students they must be prepared to defend their choices.

\[ \times = \text{problematic; } \square = \text{not problematic} \]
The design of Single-Factor, Between-Subjects Experiments

<table>
<thead>
<tr>
<th>Artifact</th>
<th>Rating</th>
<th>Artifact</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td></td>
<td>Volunteerism</td>
<td>Other Confounds</td>
</tr>
<tr>
<td>Natural History</td>
<td></td>
<td>Extraneous</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>External</td>
<td>Diffusion</td>
</tr>
<tr>
<td>Placebo Effect</td>
<td></td>
<td>Demand</td>
<td>Resentful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Characteristics</td>
<td>Demoralization</td>
</tr>
<tr>
<td>Regression to the Mean</td>
<td></td>
<td>Reactivity</td>
<td>Compensatory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Effects</td>
<td>Rivalry</td>
</tr>
<tr>
<td>Repeated Testing</td>
<td></td>
<td>Attrition</td>
<td></td>
</tr>
<tr>
<td>Instrumentation</td>
<td></td>
<td>Differential</td>
<td>Attrition</td>
</tr>
</tbody>
</table>

SAMPLE QUESTIONS (CHECK ALL OF THE OPTIONS THAT APPLY)

Q1. Which design prevents the selection artifact?
   [A] Randomized pretest-posttest control group
   [B] Nonrandomized pretest-posttest control group
   [C] Randomized posttest-only control group
   [D] None do

*Answer:* [D] (A trick question perhaps.) The selection artifact can’t be prevented. Its probability is only reduced (greatly) by randomization. Of course, it is irrelevant in single-group designs.

Q2. Which design provides protection against the occurrence of the selection artifact?
   [A] Randomized pretest-posttest control group
   [B] Randomized posttest-only control group
   [C] Randomized covariate posttest-only control group
   [D] None do

*Answer:* [A] [B] and [C] All employ randomization in the same manner. The use of a pretest or covariate is preferable since (a) they normally increase the precision of the design, (b) they allow the statistical “correction” of minor initial differences on the outcome, and (c) research consumers prefer the use of a pretest when feasible.

Q3. Suppose, in Dr. Jones’s earlier hypothetical single-group, pretest-posttest scenario, instead of providing acupuncture to all recruited participants he had recruited patients from his practice to receive acupuncture and had his friend and co-investigator concurrently enlist patients from his practice to receive a placebo procedure. Both groups’ pain was assessed at baseline and at the end of treatment. Which best describes this design?
   [A] Randomized pretest-posttest control group
   [B] Nonrandomized pretest-posttest control group
   [C] Separate cohort
   [D] None do

*Answer:* [C] The design is a separate cohort, which involves recruiting different participants for each condition.
Answer: [B] Obviously not [A], since randomization was not used. For our purposes, the separate cohort design was described in terms of different participants who were cycled through an institution or other experience during different time periods.

Q4. Now, suppose Dr. Jones used the same procedures except that he randomly assigned one of the two practices to administer the acupuncture treatment to all volunteering patients and the other practice to administer the placebo to its patients. Which best describes this design?

[A] Randomized pretest-posttest control group
[B] Nonrandomized pretest-posttest control group
[C] Randomized cluster design
[D] None do

Answer: [B], since the random assignment procedure violated the definition requiring each participant to have an equal chance to that of every other participant to be assigned to one of the two treatments. Random assignment here was basically worthless and no different from the use of two intact groups. The design would have qualified as a randomized cluster design had multiple practices been randomly assigned to administer the acupuncture vs. placebo conditions.

Q5. What is the chief advantage of employing a covariate in an experimental design?

[A] The avoidance of selection
[B] To increase statistical precision
[C] The reduced probability of producing a false positive result
[D] None of these is affected by the presence of a covariate.

Answer: [B] The presence of a covariate may help to ameliorate a minor selection artifact but it does nothing to avoid one. Given that a covariate increases statistical power (which will be expanded on in Chapter Nine), its use can reduce the probability of achieving a false negative result.

Q6. What distinguishes a randomized design from a nonrandomized (quasi-) experimental design?

[A] The quality of the inferences they produce
[B] Randomized designs are less likely to produce false positive results.
[C] Randomized designs are viewed as more credible science.
[D] None of these factors distinguish the two types of designs.

Answer: [A] [B] and [C] Randomized designs can also help prevent false negative results.

Q7. A class is administered a previously announced quiz during one session and the instructor decides to administer it during the next session, unannounced. The scores improve on the second administration. Choose the experimental that is most likely to explain this change:

[A] Regression to the mean
[B] Reactivity of the quiz
[C] Repeated testing
[D] Natural history or maturation
[E] Another experimental confound (Specify _______)

Answer: [C] Repeated testing is probably the more likely explanation. It would not be regression to the mean since the class membership wasn’t chosen on the basis of extremely low scores on the pretest. There is no reason to believe that natural history or maturation would operate over
such a brief time period, and the fact that the quiz was not a self-reported outcome variable (and there was no reason for students to suspect a second administration) would decrease the probability of a reactivity effect.

**Q8.** Which strategies are capable of reducing the probability (or extent of) differential attrition?

[A] Keeping the experimental interval brief

[B] Establishing rapport with participants

[C] Reimbursing participants for their time

[D] Employing an intent-to-treat strategy

[E] Constructing the comparison group to be more interesting or seemingly useful

[F] None of these are relevant to differential attrition.

**Answer:** [A] [B] [C] and [E]. All of these can decrease both attrition and differential attrition (the latter can only occur in the presence of the former). [E] is more directly relevant to differential attrition. (Employing an intent-to-treat strategy [D] has nothing to do with preventing either attrition or differential attrition, although it is recommended as an amelioration strategy in case either [and especially the latter] occurs.)

**Q9.** What is the primary disadvantage of randomized cluster designs as compared to most other randomized designs?

[A] They normally require more participants.

[B] They often encourage treatment diffusion.

[C] They normally experience less attrition.

[D] Randomizing clusters of participants is often more convenient.

**Answer:** [A] They require considerably more participants if the data are analyzed hierarchically (as recommended). They tend to be tactically more difficult to mount and more resource intensive. [D] can be a reasonable answer as well when considerable travel to administer the intervention and/or collect data is required of the experimental staff.

**Q10.** If two groups can be perfectly matched with respect to demographic and clinical variables, which of the following is true?

[A] There is no danger of selection.

[B] Attrition is unlikely.

[C] Randomization is unnecessary.

[D] None of these are correct.

**Answer:** [D] There may be less danger of selection, but there is always a danger of it, since the groups may differ on non-matched variables.

**Q11.** Why is the evaluation of data emanating from the extra comparison afforded by a delayed intervention often of little scientific value?

[A] It involves a single-group pretest-posttest comparison.

[B] It normally suffers greater attrition than the other comparisons.

[C] It is of key importance, since selection was controlled via the randomization process.

**Answer:** [A] and [B]. The delayed condition involves a single-group pretest-posttest design, which constitutes an extremely weak comparison, even more so here because many participants typically do not opt to receive the delayed intervention, hence it suffers from differential attrition. The delayed intervention does afford an additional comparison with the original
intervention effect, but the resulting contrast has a number of confounds (e.g., differential attri-

Q12. An educational researcher decides to test the efficacy of an instructional intervention
involving the use of manipulative aids to facilitate kindergarteners’ grasp of number con-
cepts. Two schools are selected from the same geographical area serving families with simi-
lar socioeconomic characteristics, with one being assigned to implement the intervention
and one to continue with their usual mathematical instruction. Students are taught the
same mathematical curriculum and administered the same pretests and posttests. What
“other” (i.e., other than the specific experimental artifacts introduced to this point) experi-
mental confounds might be operative here?

[A] Teacher differences
[B] Administrative differences
[C] The procedures by which students were initially assigned to teachers
[D] Time of day math classes were scheduled
[E] Class size differences
[F] These are structural differences between the schools, not experimental confounds.

Answer: [A] [B] [C] [D] and [E] All of these variables are potential confounds (i.e., have the
potential to affect the outcome variable differentially within both experimental conditions and
therefore contribute to either false positive or false negative results).

Q13. If treatment diffusion was considered a probable threat to the inferential validity of an
experiment, which design would be most likely to mitigate the potential problem?

[A] A randomized delayed-treatment design
[B] A randomized cluster design
[C] Neither would be effective in this case.

Answer: [B] Because the randomization of clusters would permit the intervention and compar-
ison groups to be physically and socially separated.

Q14. If resentful demoralization and/or compensatory rivalry were feared, which design might
mitigate the problem(s)?

[A] A randomized delayed-treatment design
[B] A randomized cluster design
[C] A patient preference design
[D] None would

Answer: [A] [B] and [C]. [A] and [C] because both participants (for resentful demoralization)
and clinicians (for compensatory rivalry) would be assured that no one who wanted the inter-
vention would be denied it. A randomized cluster design [B] would also help prevent both
artifacts since the IRB would most likely not require participants to know the identity of the
experimental condition to which they were not assigned and they would be less likely to com-
pare notes with individuals in the alternative condition.
Q15. Which condition(s) must exist for a covariate to be either appropriate or effective?

[A] A covariate must correlate at least moderately with the outcome variable in the intervention group but not the control/comparison group.

[B] A covariate must represent a variable that the intervention is capable of changing.

[C] All are contraindicated for a covariate design.

Answer: [C] It is true a covariate must correlate at least moderately with the outcome variable to be useful, but it must do so similarly within all experimental groups. Covariates must be measured (or their values must be fixed) prior to the introduction of the experimental conditions so that they cannot be affected thereby.
2 X 2 FACTORIAL DESIGNS

Crossed factorial designs testing either (a) a single intervention and one or more attribute variables or (b) two independent interventions are probably the most commonly employed factorial designs in the psychological, educational, behavioral, and health disciplines and should be given the most emphasis in this chapter. It has been my experience that the concepts involved (especially the difference between interactions and main effects and the interpretation of the former) take many students considerable amount of time to internalize.

SUGGESTED ACTIVITIES

Practice in interpreting $2 \times 2$ interactions (the most easily conceptualized) can be provided either via graphs, column and row cell subtraction, or both.

Interpreting existing data: You should feel free to substitute your own example more relevant to the students’ discipline. It can be a randomized block design involving a manipulated factor randomly assigned separately to each of two levels of an attribute variable, such as depicted here.

2 (PHONICS VS. WHOLE WORD INSTRUCTION) $\times$ 2
(ABOVE VS. BELOW AVERAGE READING ABILITY)

<table>
<thead>
<tr>
<th>RANDOMIZED BLOCK DESIGN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>First-Grade Students Reading at or Above Grade Level</td>
</tr>
<tr>
<td>First-Grade Students Reading Below Grade Level</td>
</tr>
</tbody>
</table>
Or, alternately it can involve two separate intervention factors. Let's assume the latter for our example in which an investigator wished to conduct an experiment to assess the effects of expectation on acupuncture. Let's further suppose that an easy, consistent manner of inducing temporary, nonpersistent pain at the tip of participants’ index fingers had been developed and the investigator wished to determine the role of induced expectancy played in acupuncture's ability to reduce that pain.

The two factors were (a) acupuncture vs. placebo acupuncture and (b) positively vs. negatively induced expectations. (The latter was operationally defined as the therapists assuring the participants that the procedure [acupuncture or placebo] either was known not to reduce pain or was known to be extremely effective in doing so.) Undergraduate students were accordingly randomly assigned to one of the following four cells.

### 2 × 2 FACTORIAL INVOLVING TWO INTERVENTIONS, ONE WITH A CONTROL AND ONE WITHOUT

<table>
<thead>
<tr>
<th>Positive Expectation Instructions</th>
<th>Acupuncture</th>
<th>Placebo Acupuncture</th>
</tr>
</thead>
<tbody>
<tr>
<td>a: Acupuncture with Positive Message</td>
<td>b: Placebo with Positive Message</td>
<td></td>
</tr>
<tr>
<td>Negative Expectation Instructions</td>
<td>c: Acupuncture with Negative Message</td>
<td>d: Placebo with Negative Message</td>
</tr>
</tbody>
</table>

To make the data a bit easier to interpret, let's assume that the outcome variable was expressed in terms of pain relief on a VAS scale from 0 to 10, which means that higher values reflect increased efficacy. As detailed next and mentioned in the text, there are eight possible combinations of statistically significant and nonsignificant combinations for this (or any) two-factor design. Also, listed for the 2 × 2 design described here are the 10 possible ways in which a statistically significant interaction might manifest itself.

### SUGGESTED ACTIVITY

Assign the students (in groups or individually) to construct data in the following 2 × 2 shell to reflect examples of how this experiment might produce several different statistically significant interactions with and without statistically significant main effects. (Statistical significance in this exercise can be arbitrarily defined as a one-point difference in the contrast of interest.) Alternately, the results can be graphed with other groups guessing which effects will be statistically significant and which will not. It might be wise to warn the students that it is impossible for most of the possible combinations of significant interactions in this scenario to make much
conceptual sense (so they should simply ignore reality and represent what could accrue from this strange little experiment).

2 × 2 DATA SHELL

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Expectancy Main Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectation Instructions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Expectation Instructions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acupuncture Main Effect</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 10 possible manifestations for a statistically significant interaction are illustrated here for our 2 (acupuncture vs. placebo) × 2 (high vs. low expectancy) example. We will also define statistical significance in these tables arbitrarily in terms of a one-point difference for both main effect and interaction contrasts. For pedagogical purposes, the first interaction will be graphed (Figure IM.6) and the 10 interactions (Tables IM.1–IM.10) will be illustrated by the subtraction method. The most important point to stress about a statistically significant interaction involving a treatment variable is that it potentially changes the interpretation of the treatment main effect, hence the latter (significant or not) must be interpreted on the basis of the interaction. The first two examples represent the most extreme case of this principle. In them, if the interaction is ignored, the only possible conclusion from this study is that nothing whatsoever is “going on” with these two variables while in truth a very important (even shocking) finding is “hidden” within the interaction.

Table IM.1 Placebo can be more effective than acupuncture for positive instructions, but less effective than acupuncture for negative instructions (no significant main effects)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>6.2</td>
<td>4.2</td>
<td>5.2 (6.2 + 4.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>4.2</td>
<td>6.2</td>
<td>5.2 (4.2 + 6.2)/2</td>
</tr>
<tr>
<td>Total</td>
<td>5.2 (6.2 + 4.2)/2</td>
<td>5.2 (4.2 + 6.2)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [6.2 − 4.2] vs. [4.2 − 6.2] = 2.0 vs. −2.0
In Introduction to the Experimental Process

**FIGURE IM.6:** Placebo is more effective than acupuncture for positive instructions, but less effective than acupuncture for negative instructions (no significant main effects).

**Table IM.2** Placebo can be less effective than acupuncture for positive instructions, but more effective for negative instructions (still no significant main effects)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>4.2</td>
<td>6.2</td>
<td>5.2 (4.2 + 6.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>6.2</td>
<td>4.2</td>
<td>5.2 (6.2 + 4.2)/2</td>
</tr>
<tr>
<td>Total</td>
<td>5.2 (4.2 + 6.2)/2</td>
<td>5.2 (6.2 + 4.2)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [4.2 – 6.2] vs. [6.2 – 4.2] = −2.0 vs. 2.0

**Table IM.3** Placebo can be less effective than acupuncture for both positive instructions and negative instructions, but proportionally more so for negative instructions (significant main effect for acupuncture, none for expectations)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>4.2</td>
<td>6.2</td>
<td>5.2 (4.2 + 6.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>2.4</td>
<td>6.2</td>
<td>4.3 (2.4 + 6.2)/2</td>
</tr>
<tr>
<td>Total</td>
<td>3.3 (4.2 + 2.4)/2</td>
<td>6.2 (6.2 + 6.2)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [4.2 – 6.2] vs. [2.4 – 6.2] = −2.0 vs. −3.8
Table IM.4 Placebo can be less effective than acupuncture for both positive instructions and negative instructions, but proportionally more so for positive instructions (significant main effect for acupuncture, none for expectations)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>4.2</td>
<td>6.2</td>
<td>5.2 (4.2 + 6.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>4.2</td>
<td>5.2</td>
<td>4.7 (4.2 + 5.2)/2</td>
</tr>
<tr>
<td>Total</td>
<td>4.2 (4.2 + 4.2)/2</td>
<td>5.7 (6.2 + 5.2)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [4.2 - 6.2] vs. [4.2 - 5.2] = -2.0 vs. -1.0

Table IM.5 Placebo can be more effective than acupuncture for both positive instructions and negative instructions, but proportionally more so for positive instructions (both main effects being statistically significant)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>6.2</td>
<td>2.4</td>
<td>4.3 (6.2 + 2.4)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>4.2</td>
<td>2.4</td>
<td>3.3 (4.2 + 2.4)/2</td>
</tr>
<tr>
<td>Total</td>
<td>5.2 (6.2 + 4.2)/2</td>
<td>2.4 (2.4 + 2.4)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [6.2 - 2.4] vs. [4.2 - 2.4] = 3.8 vs. 1.8

Table IM.6 Placebo can be more effective than acupuncture for both positive instructions and negative instructions, but proportionally more so for negative instructions (significant main effect for placebo not for expectations)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>6.2</td>
<td>4.2</td>
<td>5.2 (6.2 + 4.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>6.2</td>
<td>2.4</td>
<td>4.3 (6.2 + 2.4)/2</td>
</tr>
<tr>
<td>Total</td>
<td>6.2 (6.2 + 6.2)/2</td>
<td>3.3 (4.2 + 2.4)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [6.2 - 4.2] vs. [6.2 - 2.4] = 2.0 vs. 3.8

Table IM.7 No difference between acupuncture and the placebo for positive instructions, but placebo is less effective than acupuncture for negative instructions (both main effects significant)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2 (6.2 + 6.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>4.2</td>
<td>6.2</td>
<td>5.2 (4.2 + 6.2)/2</td>
</tr>
<tr>
<td>Total</td>
<td>5.2 (6.2 + 4.2)/2</td>
<td>6.2 (6.2 + 6.2)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [6.2 - 6.2] vs. [4.2 - 6.2] = 0 vs. -2.0
Table IM.8  No difference between acupuncture and placebo for positive instructions, but placebo is more effective than acupuncture for negative instructions (significant main effect for expectations, not acupuncture)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2 (6.2 + 6.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>4.2</td>
<td>2.4</td>
<td>3.3 (4.2 + 2.4)/2</td>
</tr>
<tr>
<td>Total</td>
<td>5.2 (6.2 + 4.2)/2</td>
<td>4.3 (6.2 + 2.4)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [6.2 – 6.2] vs. [4.2 – 2.4] = 0 vs. 1.8

Table IM.9  No difference between acupuncture and placebo for negative instructions, but placebo is less effective than acupuncture for positive instructions (both main effects significant)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>4.2</td>
<td>6.2</td>
<td>5.2 (4.2 + 6.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4 (2.4 + 2.4)/2</td>
</tr>
<tr>
<td>Total</td>
<td>3.3 (4.2 + 6.2)/2</td>
<td>4.3 (6.2 + 2.4)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [4.2 – 6.2] vs. [2.4 – 2.4] = −2.0 vs. 0

Table IM.10  No difference between acupuncture and placebo for negative instructions, but placebo is more effective than acupuncture for positive instructions (both main effects are significant)

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Acupuncture</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Expectations</td>
<td>6.2</td>
<td>4.2</td>
<td>5.2 (6.2 + 4.2)/2</td>
</tr>
<tr>
<td>Negative Expectations</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4 (2.4 + 2.4)/2</td>
</tr>
<tr>
<td>Total</td>
<td>6.2 (6.2 + 6.2)/2</td>
<td>5.2 (4.2 + 6.2)/2</td>
<td></td>
</tr>
</tbody>
</table>

Interaction: [6.2 – 4.2] vs. [2.4 – 2.4] = 2.0 vs. 0

A slightly different scenario: The same logic holds when the two levels of both factors contain the intervention vs. no intervention, as depicted in the next chart.
2 × 2 FACTORIAL INVOLVING TWO INTERVENTIONS, BOTH WITH CONTROLS

<table>
<thead>
<tr>
<th>Positive Expectation Instructions</th>
<th>Acupuncture</th>
<th>Placebo Acupuncture</th>
</tr>
</thead>
<tbody>
<tr>
<td>a: Acupuncture with Positive Instructions</td>
<td>b: Placebo with Positive Instructions</td>
<td></td>
</tr>
<tr>
<td>No Expectation Instructions</td>
<td>c: Acupuncture with No Instructions</td>
<td>d: Placebo with No Instructions</td>
</tr>
</tbody>
</table>

Cell a represents the combination of the two interventions, while Cell b and Cell c represent the effect of each intervention by itself. Thus these two latter cells, when compared to Cell d (the complete absence of an intervention), indicate whether acupuncture alone or positive instructions alone are superior to an appropriate control. Furthermore, if the two interventions are additive, there will be no interaction effect because Cell a – Cell b will be equivalent to Cell c – Cell d (or Cell a – Cell c will be equivalent to Cell b – Cell d).

In this scenario, the factorial design is quite advantageous because it allows the two interventions (acupuncture vs. placebo and positive expectation instructions vs. no instructions) to be tested by all of the participants instead of only half of the participants when a pairwise comparison of Cell b with Cell d and Cell c with Cell d is employed, thereby in effect doubling the sample size for the main effects. As always, however, if an interaction occurs (which it would if the combination of the two interventions was more than additive or less than additive), the main effects would not be good estimates of the effectiveness of the two interventions and the investigator would be back to analyzing the four cells via pair-wise comparisons.

SUGGESTED ACTIVITY

I include distinctions between the slightly different genres of factorial designs for two reasons:

First, it is obviously important to examine what actually comprises an experiment’s groups in order to ascertain what the experiment tests.

Second, it is important to emphasize the importance of interpreting an interaction prior to a main effect in a factorial design, otherwise it is quite possible to misinterpret the results entirely.

One way to underline these two messages is to assign groups of students to the task of designing different combinations of hypothetical 2 × 2 factorial experiments (or add 2 × 3 and 2 × 2 × 2 designs if you think the class is sufficiently advanced). Then, have the groups explain to
the class (a) what each main effect and interaction would assess and (b) whether the presence (or absence) of an interaction changed the interpretation of the effectiveness of the intervention(s).

**ACTIVITY FOR SPLIT-PLLOT DESIGNS**

Assign the task of designing a hypothetical experimental context that would require a split-plot approach. Obviously, there are many possibilities for such a design, but it might be wise to begin with the construction of a shell such as the one represented in Figure IM.7 and differentiate between the confounded and non-confounded effects.

<table>
<thead>
<tr>
<th>Plot #1</th>
<th>Plot #2</th>
<th>Plot #3</th>
<th>Plot #4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE IM.7:** Shell for a Split-Plot Design

**SAMPLE ITEMS**

**Q1.** A $4 \times 4$ factorial experiment will result in how many $p$-values for main effects and interactions?

[A] 3  
[B] 4  
[C] 16

*Answer:* [A] (Students should realize that the number of levels within a factor does not affect the number of terms.)

**Q2.** What are they?

*Answer:* The two main effects and the interaction between them.

*Questions 3–6 refer to the following chart. Assume that +/-2.0 point mean differences are statistically significant. If this involves too much calculation for your tastes (or if you feel that your students understand the effects emanating from a three-factor design), please feel free to ignore it. (Students may need a calculator here. Alternately, you might simply want to go over this or another example on the board.) I have found the calculation of mean effects and interactions from cell means such as this gives some students a better understanding of what the various terms in a factorial design mean. I’ve also found that exercises of this genre can be rather annoying for others.*
Q3. Which main effect(s) is (are) statistically significant?
[A] A
[B] B
[C] C
[D] None are

Answer: [B] and [C] because the B main effect is represented by the combined B1 vs. B2 cells or $(3 + 5 + 3 + 7)/4 - (6 + 8 + 7 + 9)/4 = 4.5 - 7.5 = -3.0$, which is greater than the posited $+/−2$ point difference. The C main effect is represented by the C1 vs. C2 contrast or $(3 + 6 + 3 + 7)/4 - (5 + 8 + 7 + 9)/4 = 4.75 - 7.25 = -2.50$, which is also greater than the $+/−2$ point boundary.

The A main effect, on the other hand, is represented by the combined A1 vs. A2 cells or $(3 + 6 + 5 + 8)/4 - (3 + 7 + 7 + 9)/4 = 5.5 - 6.5 = -1.00$, which is less than the criterion $+/−2$.

Q4. Which two-way interaction(s) is (are) statistically significant (i.e., the contrasts of interest has a difference of 2 or more)?
[A] AB
[B] BC
[C] AC
[D] None are

Answer: [D]. The most direct method of answering this question is to examine the two-way interaction means, which is done by collapsing across the ignored factor. Naturally, the computer will do this for the students with actual data, but this exercise is designed to give them a conceptual understanding of what an interaction term means. Graphing the interactions is also a good visual technique.

<table>
<thead>
<tr>
<th>(A1)</th>
<th>(C1 + C2)/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B1)</td>
<td>(a1b1) $(3+5)/2 = 4$</td>
</tr>
<tr>
<td>(B2)</td>
<td>(a1b2) $(6 + 8)/2 = 7$</td>
</tr>
<tr>
<td>(A2)</td>
<td>(a2b1) $(3 + 7)/2 = 5$</td>
</tr>
<tr>
<td>(B2)</td>
<td>(a2b2) $(7 + 9)/2 = 8$</td>
</tr>
</tbody>
</table>

Interaction Means = $(a1b1 - a1b2) - (a2b1 - a2b2) = (4 - 7) - (5 - 8) = (-3) - (-3) = 0$ NO, obviously no interaction at all.
Inroduction to the Experimental Process

BC Table

<table>
<thead>
<tr>
<th>B1</th>
<th>(A1 + A2)/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>(b1c1) (3+3)/2 = 3</td>
</tr>
<tr>
<td>C2</td>
<td>(b1c2) (5 +7)/2=6</td>
</tr>
<tr>
<td>B2</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>(b2c1) (6 +7)/2 = 6.5</td>
</tr>
<tr>
<td>C2</td>
<td>(b2c2) (8 + 9)/2 = 8.5</td>
</tr>
</tbody>
</table>

Interaction Means = (b1c1 - b1c2) – (b2c1 - b2c2) = (3 – 6) – (6.5 – 8.5) = (−3) – (−2) = −1

AC Table

<table>
<thead>
<tr>
<th>A1</th>
<th>(B1 + B2)/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>(a1c1) (3+6)/2 = 4.5</td>
</tr>
<tr>
<td>C2</td>
<td>(a1c2) (5+8)/2= 6.5</td>
</tr>
<tr>
<td>A2</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>(a2c1) (3+7)/2 = 5</td>
</tr>
<tr>
<td>C2</td>
<td>(a2c2) (7 + 9)/2 = 8</td>
</tr>
</tbody>
</table>

Interaction Means = (a1c1 – a1c2) – (a2c1 – a2c2) = (4.5 – 6.5) – (5 –8) = (−2) – (−3) = −1

Q4. Does the three-way interaction appear to be statistically significant?

[A] Yes
[B] No

Answer: [B] No. The best approach here is to create two separate graphs (see Figures IM.8 and IM.9) for one level of the analysis as illustrated in the text. In other words, the BC interaction can be graphed for the two levels of A, the AC interaction for the two levels of B, or AB on C. All are equivalent to one another. For illustrative purposes, the former will be graphed. Since the BC interaction plotted for A1 does not appear to differ dramatically from its A2 counterpart, the best guess is that there probably would not be a statistically significant interaction unless the sample size was very large. However, kudos for students who note that the C1 vs. C2 differences are greater at B1 for A2 than at either of the other three points (B2 at A2 and both B1 and B2 at A1). As always, the $p$-value associated with the ABC effect is the arbiter in an actual analysis.
Q6. How does the interpretation of a statistically significant three-group main effect differ from a statistically significant two-group main effect?

[A] The three-group main effects must be interpreted graphically, while this isn’t necessary for a two-group main effect.

[B] Both are similar in the sense that both main effects must be subjected to further analysis to determine which level(s) is (are) significantly different from which other level(s).

[C] Only the three-group main effect need be subjected to further analysis in order to determine which level(s) is (are) significantly different from which other level(s).

[D] None of these are true.

Answer: [C] If a main effect consisting of three or more groups is statistically significant, a multiple-comparison procedure needs to be conducted in order to ascertain which groups differ from one another. If the factor contains three levels, all three groups could be significantly different from one another or any combination of the three could differ from any other combination. In a two-level factor, there is only one possibility regarding which group differs from which other groups.

Q7. Which design employs a nested variable in an experiment designed to assess the effects of an intervention on individual students? (Check all that apply)

[A] Classrooms within schools are randomly assigned to intervention and control groups.

[B] Students within classrooms are randomly assigned to intervention and control groups.

[C] Neither design employs nesting.

Answer: [A], because students were not randomly assigned, but classrooms were. Therefore, students would be nested under classrooms. [B] constitutes a factorial design in which the intervention and control conditions are crossed with classrooms.

Q8. What is the primary tradeoff associated with a two-factor, split-plot design employing two interventions as opposed to a crossed factorial model?

[A] Both of the main effects are confounded with the plots.

[B] One of the main effects is confounded with the plots.

[C] There are no tradeoffs if the plots can be randomly assigned to participants.

Answer: [B] Only one of the two main effects need be confounded with the plots.

Q9. What is (are) the primary difference(s) between the plots used in a split-plot design and the institutions used in a randomized cluster design?

[A] Institutions cannot be used as plots.
InTroducTIon To THE EXPERIMEnT AL ProcEss

[B] Participants are randomly assigned to plots but can’t be randomly assigned to clusters.
[C] At least one factor must be confounded with plots but not with clusters.
[D] Randomization cannot be used in split-plot designs.
[E] None of these differentiate the two designs.

Answer: [C] In a typical randomized cluster design, participants are not randomly assigned to experimental groups. Instead, clusters (e.g., institution, classroom) are randomly assigned to groups, hence participants are nested within clusters. If multiple factors (such as the addition of an attribute variable [e.g., gender or severity of illness]) is added to (i.e., in addition to) experimental conditions in a randomized cluster design, the former is not confounded with the clusters. In a split-plot design, however, the plots themselves are associated with differences on an independent variable, which means that that variable is completely confounded with potential differences between plots.

Q10. How would participants be optimally assigned in a 2 (male vs. female) × 2 (experimental vs. control) factorial design?

[A] Males and females would be randomly assigned to each experimental condition.
[B] Experimental conditions would be randomly assigned to males and females separately.
[C] Since participants cannot be randomly assigned to a gender main effect, some sort of matching or split-plot design would be required.

Answer: [B] Males and females would be assigned separately to the two experimental conditions.

Q11. Which of the following characterize a fractional factorial design?

[A] It does not require the investigator to assign participants to all of the treatment combinations, thereby decreasing the required sample size.
[B] It allows participants to be assigned to multiple interventions, thereby decreasing the required sample size.
[C] It allows the investigator to ignore those interventions that are of no interest.

Answer: [A] A one-half fractional factorial requires participants to be assigned to only one-half of the total number of treatment combinations available. Not [B], because it is a between-subject design, which means that a single participant can only be assigned to one treatment combination. Not [C], because participants will be assigned to all of the interventions, just not all of the combinations thereof. Also, the investigator is constrained with respect to the combinations he or she must include and cannot pick and choose.

Q12. What is the primary disadvantage of fractional factorial designs?

[A] They can’t involve control groups, only active interventions.
[B] They are incapable of providing interpretable main effects.
[C] Some of their effects are always confounded with one other.

Answer: [C] The highest level interaction cannot be computed and the next two levels are confounded with one another. Thus, if the experiment has five factors, the five-way interaction will not be computed and the four-way and three-way interactions will be confounded with one another. However, the two-way interactions and the main effects are interpretable. If the experiment has only three factors, the three-way interaction cannot be calculated and the two-way interactions and main effects are confounded with one another, hence the strategy normally makes sense only for experiments involving five or more factors.
Repeated measures (also called within-subjects) designs can be extremely powerful experimental options when it is feasible to introduce the experimental conditions to the same or matched participants. Two powerful experimental strategies (supplementary to randomization) and especially essential (but not limited) to repeated measures designs were introduced (and without which repeated measures designs would not be nearly as useful):

1. Counterbalancing (which controls for order effects and aids in the avoidance or facilitates the detection of carryover effects), and
2. Blocking, which has been discussed previously and involves the purposeful grouping of experimental units, often to facilitate counterbalancing or to simply randomly assign participants with them, for the express purpose of eliminating (or controlling) extraneous sources of outcome variation.

These two concepts were important enough to merit their own experimental principle:

**PRINCIPLE #20:** In addition to randomly assigning participants, counterbalance all procedural components employed in an experiment via the use of blocks and then randomly assign participants to those blocks to avoid confounding.

This principle ranks among the most important design (as opposed to acculturation-related) dicta presented in the text and could have easily been presented earlier. Its formal introduction was delayed in order for it to be buttressed by an appreciation of the marriage of blocking and counterbalancing as they occur in repeated designs. If possible, this 20th principle should be mentioned frequently when students are given an assignment to design a study or to evaluate one already conducted.
DESIGNS INTRODUCED IN CHAPTER EIGHT

The two most commonly employed repeated measures/within-subjects designs introduced in this chapter were crossover and randomized matched-block designs. The former typically employs both counterbalancing and blocking while the latter uses a rather unique application of blocking. More emphasis is placed on crossover designs here, but the randomized matched-block design is, in my opinion, an underutilized, undervalued, and extremely powerful strategy when all participants can be identified at the onset of an experiment.

Quasi-experimental interrupted time series designs: Although more commonly encountered in retrospective studies, these repeated measures designs are also relevant to prospective experimentation involving human participants. They are less acceptable strategies than both counterbalanced repeated measures designs and randomized matched-block designs and are basically not recommended strategies for most prospective purposes. As always, there are exceptions to all such strictures, however, as witnessed by the very rarely employed multiple site, phased-intervention time series design, which is probably an adequate design employing nonreactive outcome variables and situations in which a placebo or Hawthorne-like effect are not plausibly occurring experimental artifacts. Should you consider one of the other designs discussed in this chapter potentially useful to your students, by all means supplement the text’s discussion with an example from your discipline. Other repeated designs discussed were (a) multiple group nonrandomized time series designs, (b) single-group designs in which the intervention and control group are implemented at least twice, and (c) the single-group interrupted time series design (analogous to the multiple-group time series design but without the additional control afforded by a comparable, preferably randomly assigned).

SUGGESTED ACTIVITY

Crossover designs are most commonly associated with counterbalancing blocks of participants comprising different orders in which the experimental conditions are run. Naturally, additional examples of actual studies in the students’ disciplines would be helpful. If none come to mind, useful strategies might be to have students (as always individually or in groups) locate such examples. Alternately, students might design their own brief scenarios employing a crossover design (or one of the other repeated measures designs introduced in the chapter) which would then be critiqued either by you or the class as a whole. As an example of instructions that might be given for this activity employing a crossover design:

1. Design a crossover experiment that might be used in your area of concentration. Accompany it with a formal hypothesis, which should encompass the five constituents recommended in Chapter Three (the intervention and its duration/dose, the comparison, the outcome, type of participants, and expected results).

2. Draw a figure to represent your design and describe the procedures involving the use of counterbalancing and blocking in your design.

Let’s suppose the following scenario represented one of the less than sterling efforts resulting from this assignment:
The experiment involved investigating the efficacy of underlining key elements of a prose passage as compared to simply reading without underlining. The outcome was free recall of the passage's factual content. On the first day of the experiment half of the class was randomly assigned to read the passage using underlining while the other half of the class was instructed to simply read and reread the passage for a given period of time. Both groups were given the same amount of study time and then asked to write down all of the facts they could remember from the passage. The next day, the students who had underlined the passage on the first day read and reread the passage while the students who read and reread the passage the first day underlined it the second day. Both groups were again tested on their recall of factual information. A repeated measures, counterbalancing design was employed because the students who were assigned to the underlining condition on the first day were assigned to read and reread the passage without underlining on the second day. The opposite order occurred for the other half of the student. The order in which the students were assigned to the experimental conditions constituted a blocking variable.

*Hypothesis:* Using a two-session crossover design, undergraduates will recall more of a prose passage when they underline key facts in the passage than when they read and reread the same passage for the same length of time without underlining.

Diagrammatically the design might be depicted as follows.

**AN EXAMPLE OF A STUDENT’S HYPOTHETICAL (AND UNACCEPTABLE) CROSSOVER DESIGN**

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order #1</td>
<td>Underlining</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td>Read-reread</td>
<td>Underlining</td>
</tr>
<tr>
<td>Order #2</td>
<td>Read-reread</td>
<td>Test</td>
</tr>
</tbody>
</table>

Now obviously this experiment involves a rather bizarre strategy, but don’t be surprised if you get something comparable. You might even consider providing the class with an example such as this for pedagogical purposes. Depending on your personality, you could pretend that it (or something similar) was a study that you just completed and were quite proud of (or alternatively you could pretend that a colleague had just done so.) Hopefully, some of the following objections would be raised.

1. A crossover design is not appropriate for this experiment because the outcome variable will not return to baseline. True, some students will forget much of what they read the first day, but everyone will remember something. Therefore, overall performance for both groups will be better on Day 2 than it was on Day 1.
2. If underlining was actually a superior learning strategy, then the students who were assigned to read and reread the passage on the second day would start out with an advantage on Day 2, having underlined the passage the previous day, while their Day 2 counterparts had not been afforded that advantage. However, the Day 2 underlining students would have an advantage over the students who underlined on Day 1, since they had already studied the passage. In any case, this study is a recipe for disaster and
would almost certainly result in a significant main effect for “days” and a day × experimental condition interaction.

Now hopefully no one will come up with a design such as this, but some will select transitory interventions with outcomes that can be counted on to revert quickly back to baseline. However, if something similar were to be presented, a reasonable question to facilitate discussion might simply be: “How could this experiment be improved?”

The easiest (and most direct) answer would probably be to abandon a crossover design altogether and employ one of the between-subjects designs discussed in the previous chapter. Since a pretest wouldn't make much sense here (and a covariate might be hard to come by for undergraduates who are probably quite homogeneous with respect to reading ability), the best candidate for this particular study might be a randomized posttest-only control group design.

**SAMPLE ITEMS**

*Scenario for Questions 1–5:* A well-designed two-group crossover design was employed to ascertain if feedback regarding performance on a task resulted in fewer errors than no feedback. Students were randomly assigned to receive the order in which they received the feedback intervention as follows:

```
Order #1  O₁ Feedback  O₂ O₃ Control  O₄
Order #2  O₁ Control   O₂ O₃ Feedback O₄
```

Let’s assume for analytic purposes, after the data were collected the experiment was reconfigured as follows:

```
Order #1  O₁ Feedback  O₂ O₃ Control O₄
Order #2  O₁ Feedback  O₄ O₂ Control O₃
```

**Q1.** What would comprise an optimal outcome for this experiment?

[A] No statistically significant order effect, but a significant interaction

[B] Both the order and the treatment effect being statistically significant but no statistically significant interaction

[C] All three effects (order, treatment, and the interaction between them) reaching statistical significance

[D] Only the treatment main effect being statistically significant

[E] None are optimal.

*Answer:* [D] Either a statistically significant (a) order main effect or (b) order by treatment interaction would be difficult to explain in a crossover experiment such as this (assuming the data were reconfigured as indicated).

**Q2.** What could a statistically significant order × treatment interaction indicate in this design?

[A] The feedback was only effective for one group (order) and not the other.

[B] The control was superior to feedback in one group but inferior in the other.
Repeated Measures, Within-Subjects, and Longitudinal Designs

[C] The feedback was superior to the control in one group but not different from the control in the other.
[D] Feedback was superior to the control in both groups but more so in one than the other.
[E] None of these

Answer: [A], [B], [C], and [D] All of these options could result in a statistically significant interaction. A problem with the crossover design is that when an order effect or interaction occurs, its etiology is difficult (and sometimes impossible) to tease out.

Q3. Suppose, in the rudimentary crossover design presented here, the author found feedback to produce significantly fewer errors than control in Order #1 but only a very small, nonsignificant reduction in errors in Order #2. What could explain such a finding (assuming that the interaction effect was statistically significant)?

[A] Once participants had experienced feedback, they came to rely on it and this depressed performance when no feedback was present.
[B] Participants became fatigued during the course of the experiment and tended to commit more errors on the second presentation of the task.
[C] Participants originally found the task interesting and performed it conscientiously, but became bored with it over time.
[D] None of these could explain why the intervention failed to produce fewer errors in both orders.

Answer: [B] and [C]. Both explanations are consistent with the superiority of the intervention in Order #1 and not Order #2. In other words, the novelty of the task (or alertness of the participants) enhanced performance for the feedback group in Order #1 and the control in Order #2. The finding that the intervention was slightly better than the control in Order #2 is irrelevant because, by the rules experimenters play by, if a finding is not statistically significant it is considered not reliable. ([A] does not explain why feedback was not effective for Order #2.)

Q4. Given the results posited in Q3, what would a reasonable course of action be for the investigator who obtained them if he or she remained convinced that feedback was a viable intervention?

[A] Simply not attempt to publish the results
[B] Publish the Order #1 results, which demonstrated a statistically significant effect for feedback
[C] Attempt to publish the results as obtained and offer the most plausible alternative explanations possible for the failure of Order #2 to support the hypothesis
[D] Redesign the study to address as many alternative explanations for the lack of effect for Order #2 and include both experiments in the same paper

Answer: [C] is reasonable, but does not address the issue and makes a very minor contribution to the literature. [D] is preferable because regardless of the results (i.e., whether feedback is found to be superior or not), more light will be shed on the conditions under which feedback does or does not facilitate performance. [A] is a waste of the participants’ efforts and the other resources consumed in conducting the experiment. [B] is contraindicated for two reasons: first, it violates Principle #1 because it is dishonest and disingenuous; second, a one-group crossover design constitutes an extremely poor experimental choice and is little if any better than no experiment at all.
Q5. What would constitute a reasonable redesign of this experiment given the equivocal results posited in Q3? This one is a little involved and may require being drawn on the board. Alternately, it could serve as an exercise in which students redesign the experiment. (Even if crossover designs are not a high priority in your discipline, exercises such as this may help students think through the implications and alternative explanations for design decisions in general. Naturally, I have no evidence, so as always, use your own judgment as to whether to go over this question in class.)

[A] Employ two different tasks via the following design:

<table>
<thead>
<tr>
<th>Task A</th>
<th>Task B</th>
</tr>
</thead>
<tbody>
<tr>
<td>O₁, O₂</td>
<td>Feedback</td>
</tr>
<tr>
<td>O₃, O₄</td>
<td>Control</td>
</tr>
<tr>
<td>O₅, O₆</td>
<td></td>
</tr>
</tbody>
</table>

[B] Employ two different tasks on two different days via the following design:

<table>
<thead>
<tr>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task A</td>
<td>Task B</td>
</tr>
<tr>
<td>O₁, O₂</td>
<td>Feedback</td>
</tr>
<tr>
<td>O₃, O₄</td>
<td>Control</td>
</tr>
<tr>
<td>O₅, O₆</td>
<td>Feedback</td>
</tr>
<tr>
<td>O₇, O₈</td>
<td>Control</td>
</tr>
<tr>
<td>O₉, O₁₀</td>
<td>Feedback</td>
</tr>
<tr>
<td>O₁₁, O₁₂</td>
<td></td>
</tr>
</tbody>
</table>

[C] Simply replicate the original design on the second day with a different task:

<table>
<thead>
<tr>
<th>Day 1 (Task A)</th>
<th>Day 2 (Task B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order #1</td>
<td>Order #2</td>
</tr>
<tr>
<td>O₁, O₂ Feedback</td>
<td>O₁, O₂ Control</td>
</tr>
<tr>
<td>O₃, O₄ Control</td>
<td>O₅, O₆ Feedback</td>
</tr>
<tr>
<td>O₇, O₈ Control</td>
<td>O₉, O₁₀ Feedback</td>
</tr>
<tr>
<td>O₁₁, O₁₂</td>
<td></td>
</tr>
</tbody>
</table>

[D] Replicate the original design on Day 2 but counterbalance the order of presentation employed on Day 1

<table>
<thead>
<tr>
<th>Day 1 (Task A)</th>
<th>Day 2 (Task B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order #1</td>
<td>Order #2</td>
</tr>
<tr>
<td>O₁, O₂ Feedback</td>
<td>O₁, O₂ Control</td>
</tr>
<tr>
<td>O₃, O₄ Control</td>
<td>O₅, O₆ Feedback</td>
</tr>
<tr>
<td>O₇, O₈ Control</td>
<td>O₉, O₁₀ Feedback</td>
</tr>
<tr>
<td>O₁₁, O₁₂</td>
<td></td>
</tr>
</tbody>
</table>

Answer: [D] and to a lesser extent [C]. [D] would probably find greater favor with journal reviewers since they counterbalances the order of intervention presentation for all participants. It isn’t perfect, of course, because the two days are confounded with the two tasks and there could quite possibly be a main effect for the day/task main effect. (This would be irritating but not fatal if the feedback vs. control differences were relatively constant across both days/tasks and both orders. Said another way, if the final data were reconfigured and there was a day/task main effect but no day/task × treatment interaction, then very good evidence would be
produced for the efficacy of feedback for this type of performance. Reconfigured, this design would look like Figure IM.10.

**Figure IM.10: Reconfigured 2 (Conditions) × 2 (Task) × 2 (Day) Crossover Design**

<table>
<thead>
<tr>
<th></th>
<th>Feedback</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td>Day 2</td>
<td>Day 1</td>
</tr>
<tr>
<td>Order #1</td>
<td>Day 1/Task A</td>
<td>Day 2/Task B</td>
</tr>
<tr>
<td>Order #2</td>
<td>Day 1/Task A</td>
<td>Day 2/Task B</td>
</tr>
</tbody>
</table>

If no order or day main effects surfaced and no interactions involving the two with either experimental conditions or type of tasks, the design depicted in Figure IM.10 could be reconfigured conceptually as a 2 (Task A vs. Task B) × 2 (Feedback × Control) design where both factors are repeated measures:

<table>
<thead>
<tr>
<th>Task A</th>
<th>Task B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback</td>
<td>Control</td>
</tr>
</tbody>
</table>

And if there were no task main effect or task × condition interaction, the entire design conceptually (obviously it wouldn’t be reported this way) reduces to a two-group repeated measures configuration:

| Feedback | Control | Feedback | Control |

Design [A] doesn’t “reduce” to a repeated measures design but must rely on a between-subjects contrast for the feedback vs. control condition for each task separately. Design [B] is really no better; it never reduces to a repeated measures feedback vs. control contrast involving the same task.

**Q6.** Which of the following is true of a randomized matched-block design?

- [A] The same participants are administered multiple experimental conditions.
- [B] All study participants must be recruited before the administration of the experimental conditions.
- [C] The blocking variable must not be empirically or conceptually related to the outcome variable.
- [D] The blocking variable must be empirically related to the outcome variable in order to be effective.
- [E] None of these are characteristics of a randomized matched-block design

**Answer:** [B] [D] The major drawbacks of the design are that participants must be available prior to randomization and a blocking variable correlated with the outcome variable must be identified and measured prior to the administration of the treatment. This blocking variable should also be continuous to avoid ties.
Q7. Which design(s) is (are) superior to the single-group pretest-posttest design?

**Answer:** [A] [B] and [C] All are because of the multiple administrations of the outcome variable. Even [A] is superior because of at least some increased protection against the *post hoc, ergo propter hoc* logical fallacy. Also, several experimental artifacts are vitiated (e.g., external extraneous events, regression to the mean, natural history or maturation) if an effect occurs at the predicted measurement (i.e., most likely between $O_8$ and $O_9$).

Q8. What information do the multiple measurements following the intervention’s implementation supply besides the efficacy of the intervention?

**Answer:** Assuming there is an intervention effect, they provide information about its rate of decay (or staying power).

Q9. Assuming statistically significant positive intervention results at or near the predicted post-intervention assessment, which design would provide the strongest causal efficacy conclusion?

**Answer:** [C] in the absence of an artifact that requires an appropriate control group, such as the possibility of a placebo or Hawthorne-type effect.

Q10. Which design provides the most protection against selection?

[A] Randomized matched block  
[B] Randomized pretest-posttest control group  
[C] Randomized posttest-only control group  
[D] A nonrandomized design that selected individuals from the same institution and carefully matched them on multiple variables including the pretest

**Answer:** [A], because the matching procedure, followed by randomizing participants to groups provided even more protection than [B] or [C], both of which largely negate the threat. [D]
Repeated Measures, Within-Subjects, and Longitudinal Designs

represents an improvement over a nonequivalent pretest-posttest control group design. (It is extremely difficult, however, to match participants on multiple variables unless they are combined into a propensity score.)

Q11. What is (are) the primary advantage(s) of a Latin square design?
[A] It ensures that each experimental condition both precedes and follows all other conditions.
[B] It ensures that each experimental condition immediately (i.e., adjacently) precedes and follows each other condition.
[C] It ensures that each experimental condition occupies each possible order of presentation in the experiment.

[D] None apply

Answer: [A] and [C]. The Latin square design does not guarantee that each condition will immediately precede and follow each other condition, only that everything will precede and follow everything else and occupy each position (e.g., 1st, 2nd, 3rd, and 4th for a four-group square).

Q12. Which repeated measures design might be used when too many experimental conditions exist to reasonably be completed by all participants?
[A] A counterbalanced crossover design
[B] A non-counterbalanced crossover design
[C] A single-group time series design
[D] A balanced incomplete block design
[E] None would solve the problem.

Answer: [D], since each respondent is administered only a subset of the total number of experimental conditions. None of the other options address the problem.
The concept of statistical power is difficult for most students. To help clarify the issue I would suggest that you use the following definition rather than some of the more succinct options available in the literature because it stresses the importance of the design, conduct, and predicted effect size of the experiment:

Statistical power is defined as the probability that an experiment will result in statistical significance if that experiment is properly designed and properly conducted and its hypothesized effect size is correct.

I’ve also found it helpful to differentiate statistical significance from statistical power even if the distinction seems obvious.

Statistical significance is computed after a study is completed and its data are collected and analyzed. It is used to estimate how probable the study’s obtained differences or relationships (both of which are called effect sizes) would be to occur by chance alone.

Statistical power, on the other hand, is computed before an experiment’s final data are collected via a two-step process involving (1) hypothesizing both the direction and the size of the effect most likely to occur as a result of the intervention and (2) estimating how probable the study’s results are to result in statistical significance if both the direction and the size of the effect were correctly specified.
The importance of the process is underlined by the addition of (a) a 17th inferential artifact accompanied by (b) two experimental principles designed to avoid it:

**Experimental Artifact #17: Underpowered experiments**: defined as experiments which have been improperly designed to avoid producing a false negative result.

**PRINCIPLE #21**: Always conduct a power/sample size analysis as part of the design process and ensure that the experiment has sufficient power to detect the hypothesized effect size.

**PRINCIPLE #22**: Do not subvert the power/sample size analytic process by making it a process for justifying a convenient, feasible, or easily available sample size.

### ACTIVITY/DISCUSSION

The following chart, reproduced from Chapter Nine, is often used in discussions of statistical power. It has been my experience that some students find it helpful while others find it confusing. I have included some sample items to facilitate mastery of the basic concepts, but if you decide to use them you should probably introduce it beforehand.

<table>
<thead>
<tr>
<th>Possible Experimental Outcomes</th>
<th>The Hypothesis Is Correct</th>
<th>The Hypothesis Is Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Hypothesis Is Confirmed</strong></td>
<td>[a] Correct ( (p \text{ of occurrence}) = \text{statistical power} = .80 )</td>
<td>[b] Incorrect ( \text{[Type I Error]} ) ( (p \text{ of occurrence}) = .05 )</td>
</tr>
<tr>
<td>( p &lt; .05; \text{ power set at .80} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The Hypothesis Is Not Confirmed</strong></td>
<td>[c] Incorrect ( \text{[Type II Error]} ) ( (p \text{ of occurrence}) = 1 - \text{statistical power} = .20 )</td>
<td>[d] Correct ( (p \text{ of occurrence}) = 1 - \alpha = .95 )</td>
</tr>
<tr>
<td>( p &gt; .05; \text{ power set at .80} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It may also be worth mentioning that the concepts represented in this table and the strategies offered in this chapter are all designed not only to avoid designing an underpowered experiment (Experimental Artifact #17) but also to avoid wasting resources by recruiting and running too many participants.

### THE EFFECT SIZE

On one level the ES is a mathematically simple concept (especially Cohen's \( d \), which is the only one presented), but its specification prior to conducting an experiment is counterintuitive to many students. I offer only two formulas in this book (both in the present chapter) with the ES \( (d) \) being the only one I feel is important for students to memorized, internalize, and have a bit of practice calculating.
STRATEGIES FOR INCREASING STATISTICAL POWER

The 12 strategies for increasing statistical power which students should be aware of are as follows:

**Strategy #1:** Increasing the sample size.

**Strategy #2:** Assigning more participants to groups (often control groups) which are less resource intensive to run.

**Strategy #3:** Using as few groups as possible.

**Strategy #4:** Increasing the size of the hypothesized ES.

**Strategy #5:** Employing covariates and/or blocking variable.

**Strategy #6:** Employing crossover or repeated measures designs.

**Strategy #7:** Hypothesizing between-subjects main effects rather than interactions.

**Strategy #8:** Randomizing individual participants rather than groups of participants.

**Strategy #9:** Employing measures sensitive to change.

**Strategy #10:** Employing reliable outcome variables.

**Strategy #11:** Using proximal as opposed to distal outcome variables.

**Strategy #12:** Using appropriate data analysis procedures.

It is important to stress the fact that the first 11 all involve tradeoffs of one sort or another and these tradeoffs must be evaluated on the basis of scientific considerations, which trump all other concerns. (The 12th is a scientific necessity.) It is also important to stress that each of these 11 strategies should be considered at the design stage of any experiment, because all have the potential of increasing the sensitivity and validity of the hypothesis test itself. Furthermore, it is difficult to overemphasize the importance of maximizing the sensitivity of one’s experiments, which all 12 of the above principles help to ensure in one way or another.

It may also be helpful to inform students that a power/sample size analysis has become an absolute necessity for almost all IRB and grant submissions (not to mention a prerequisite for publishing one’s experiments in many journals). To further emphasize the importance of the process, you might wish to supply students with the following checklist to be completed at the design stage of the studies (this checklist was not included in the text and is presented here for pedagogical purposes).

1. Is the sample adequate and properly justified?
2. Is the hypothesized ES realistic and empirically* justified?
3. Have covariates and/or blocking variables been identified and incorporated?
4. Have the adjustments to the hypothesized ES been empirically justified?
5. Could the intervention be administered as a repeated measures design?
6. If so, was the ES appropriately adjusted (and empirically justified)?
7. If the hypothesis involved an interaction, was the ES adjusted appropriately?
8. If clusters were randomized, was the ES appropriately adjusted via their ICC?

* By “empirically justified” I mean via a pilot study, a meta-analysis, or similar experiment in the literature. I suppose we should also include something like Jacob Cohen’s small, medium, and large rationale for ESs.
9. Was this ICC empirically justified?
10. Has the outcome variable been shown to be sensitive to change in other studies?
11. Is there sufficient theoretical justification for the outcome variable’s proximal relationship to the intervention?

SAMPLE ITEMS

Questions 1–4 refer to the following figure (Figure 9.1 in the text):

**Figure 9.1 Four Experimental Outcomes and Their Probabilities**

<table>
<thead>
<tr>
<th>Possible Experimental Outcomes</th>
<th>What Is Actually True</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The Hypothesis Is Correct</td>
</tr>
<tr>
<td>The Hypothesis Is Confirmed</td>
<td>[a] Correct [(p of occurrence) = statistical power = .80]]</td>
</tr>
<tr>
<td>(p &lt; .05; \text{power set at .80})</td>
<td></td>
</tr>
<tr>
<td>The Hypothesis Is Not Confirmed</td>
<td>[c] Incorrect [Type II Error] [(p of occurrence) = 1 – statistical power = .20]]</td>
</tr>
<tr>
<td>(p &gt; .05; \text{power set at .80})</td>
<td></td>
</tr>
</tbody>
</table>

Q1. Once the experiment is run and statistical significance is obtained, which cells(s) can the investigator be absolute certain was (were) correct?
   - [A] Cell a
   - [B] Cell b
   - [C] Cell c
   - [D] Cell d
   - [E] None

   **Answer:** [E] This and the next question highlight the limitation of this chart and its theoretical nature. Either before or after an experiment is conducted, we never know what the *truth* is (i.e., whether the hypothesis is actually correct or incorrect). This chart is therefore designed to ascertain a priori probabilities in the presence of ignorance. If statistical significance is obtained after the experiment is conducted, we do know that a Type II error did not occur, but we still do not know that the hypothesis was actually correct, thus we may or may not have committed a Type I error.

Q2. If statistical significance had not been obtained?
   - [A] Cell a
   - [B] Cell b
   - [C] Cell c
   - [D] Cell d
   - [E] None

   **Answer:** [E], for the same reason, except this time we know that a Type I error did not occur.
Ensuring Sufficient Statistical Power

Q3. If the investigator had set the alpha level at .01 and accordingly designed the experiment to have power of .80, what would his or her a priori probability of obtaining a false positive result?

[A] .95  
[B] .20  
[C] .01  
[D] .99  
[E] Impossible to ascertain

*Answer:* [C] Since this deals with the a priori probability of obtaining a false positive result (which is called a Type 1 error), Cell b applies, which is the alpha level, or .01.

Q4. Of obtaining a false negative result?

[A] .80  
[B] .20  
[C] .01  
[D] .99  
[E] Impossible to ascertain

*Answer:* [B] The experiment is still powered at .80, hence Cell d applies (1 – power = 1 – .80 = .20).

Q5. Which of the following is or are characteristic of an ES?

[A] It is independent of the scale on which the outcome variable is measured.  
[B] It is a measure of the strength of an effect and is therefore independent of the normal curve.  
[C] It is independent of the characteristics of the sample.  
[D] None of these

*Answer:* [A] The ES is a standardized measure expressed in terms standard deviation units and therefore is scale nonspecific (e.g., the outcome values can be expressed in terms of something like IQ with a mean of a 100 or a 11-point VAS, but an ES of 1.0 still indicates that the intervention and the control differed by exactly one standard deviation on both measures). The ES is not independent of differences in samples, however, since homogeneous study participants will have a smaller standard deviation than a heterogeneous sample. And anything based on a standard deviation is based on the normal curve.

Q6. Name the three numerical values needed to determine the statistical power available for a two-group (between subject) experiment.

*Answer:*

1. The alpha level (significance level) that will be used to test the hypothesis
2. The number of participants to be assigned to each group
3. The hypothesized ES

Q7. What three numerical values are needed to determine the required sample size for such an experiment?

*Answer:*

1. The alpha level (significance level) that will be used to test the hypothesis
2. The desired power level
3. The hypothesized ES

Q8. Assuming that a pilot study was conducted to estimate the required sample size for the final experiment, which of the following results would require the larger sample size assuming the pilot results were valid?

[A] Intervention (Mean 14; SD 4) vs. Control (Mean 12, SD 4)
[B] Intervention (Mean 14; SD 10) vs. Control (Mean 10, SD 10)
[C] Impossible to tell

Answer: [B] The point this question is trying to emphasize is the role of the ES in the determination of an optimal sample size. Since the ES is calculated by dividing the mean difference between two groups by the pooled standard deviation, the ES for [A] is 14 − 12/4, or 0.50; for [B] it is 14 − 10/10 = 0.40—which is the correct answer since the smaller the predicted ES, the larger the required N/group to achieve the same degree of power.

Q9. Suppose our Q8 investigator was now faced with the task of choosing an alpha level for the impending experiment. Which of the following alpha levels would require the smallest sample size, assuming that a one-tailed test could be justified?

[A] .01 with a two-tailed test
[B] .01 with a one-tailed test
[C] .05 with a two-tailed test
[D] Impossible to tell

Answer: [C] , because the higher the a priori alpha level is set, the more power is available, assuming no other parameter is changed. Thus [B] would require a smaller sample size than [A] because a one-tailed alpha of .01 is equivalent to a two-tailed alpha of .02 with respect to the a priori alpha level. But obviously, [C] is higher than .02.

Q10. Which of the following strategies are likely to increase an experiment’s effect size?

[A] Doubling the sample size
[B] Increasing the sample size by randomly assigning more participants to one group than to the other
[C] Employing a no-treatment control instead of an alternative-treatment control group
[D] Increasing the strength of the intervention
[E] None of the above

Answer: [C] and [D] Increasing or decreasing the sample size affects the power available for an experiment but does not itself affect the ES, although it increases the statistical power available for the experiment. Employing a no-treatment control will most likely result in a larger ES than that with an alternative treatment since the latter would be expected to produce at least some outcome changes (and since the numerator of the ES formula entails subtracting the mean of the control/comparison group from the intervention mean, the smaller the numerator becomes the smaller the ES becomes). Increasing the strength of the intervention will have the opposite effect on the numerator and thus increase the ES.

Q11. Which relationship will have the greatest effect on the statistical power of an experiment?

[A] An r of 0.50 for the covariate across experimental conditions in a between-subjects design
Ensuring Sufficient Statistical Power

[B] An r of 0.50 for the outcome variables correlation across experimental conditions in a repeated measures design

[C] A reliability of 0.70 for the outcome variable across experimental conditions

[D] None of these coefficients will affect the power of an experiment.

Answer: [B] Repeated measures designs are more powerful than between-subjects designs employing a covariate. Hence, if the r between the covariate and the outcome is not considerably higher than the r between the outcome changes in a repeated measures design, then the repeated measures design will be more sensitive. Reliability is only a factor in a power analysis for experimental research if it is low or dramatically different than expected. Measurement theory hasn’t been covered in this text, but a moderate reliability coefficient of 0.70 would preclude either contingency.

Q12. Given a moderately sized ICC, which strategy will have the most influence on the total sample size required for a randomized cluster design?

[A] The number of clusters assigned

[B] The number of participants present per cluster

[C] Both are equally important as long as the total number of participants is the same.

Answer: [A] In the presence of the same ICC (20 clusters of 10 participants each produces more power than 10 clusters of 20 participants each) unless the ICC is zero.

Q13. Why is Principle #22 (not subverting the power/sample size analytic process by making it a process for justifying a convenient, feasible, or easily available sample size) in the best interests of an investigator?

[A] The violation of this principle involves scientific misconduct.

[B] The principle’s violation increases the probability of a Type II error.

[C] Violation of the principle increases the likelihood that the investigator’s hypothesis will not be supported (assuming he or she posited a difference between groups).

Answer: [A] [B] and [C] Anytime a scientist knowingly influences the outcome of an analysis or is purposefully nontransparent, he or she has committed scientific misconduct. (It may not be a censurable offense, but it still violates our first principle.) For example, forcing the power analysis to produce an acceptable level of power (usually 0.80) for a fixed sample size (e.g., by overestimating the most likely ES to accrue) increases the probability of obtaining a Type II error. And, except in the rare instance of a noninferiority hypothesis, this means that the investigator’s hypothesis is less likely to be supported.

Q14. The proper performance of a power analysis and a statistical analysis are both necessary for ensuring what kind of validity?

[A] Internal validity

[B] Statistical conclusion validity

[C] External validity

Answer: [B] You may have noticed that I have been reluctant to stress the memorization of terms, since I’m far more concerned with imparting the understanding of concepts and good scientific practice. I do think that students should at least be conversant with these three terms simply because they are frequently used in some areas (although not in others) and scientists should be able to communicate effectively with one another across disciplines. The term internal validity (which
In this chapter another form of experimental validity, called by some statistical conclusion validity, has been introduced. Statistical conclusion validity refers to the process of ensuring that the statistical analysis of an experiment’s data (including the initial power analysis) is performed appropriately and thereby also contributes to the production of an incorrect experimental inference. In Chapter Eleven the third experimental validity concept (external validity or generalizability) will be discussed in considerable detail.

**Q15.** What is the primary difference between a noninferiority experiment and a conventional efficacy study?

[A] A larger sample size will normally be required.

[B] The expected ES is not calculated in the same manner.

[C] The logic behind Type I and Type II errors changes.

[D] None of these factors are impacted.

[A] [B] and [C] In a conventional experiment the larger the sample size and the larger the expected ES, the more likely it is that the experimental hypothesis will be accepted. In an experiment in which no differences are expected or hypothesized, smaller sample sizes and smaller ESs increase the likelihood of the hypothesis’ acceptance. Thus the logic of Type I and Type II error rates is turned upside down.
This chapter begins with the pejorative statement that the most elegantly designed, but sloppily conducted, experiment is worthless, which in turn introduces the book’s final experimental artifact—the avoidance of which constitutes the entire subject matter of this chapter.

Experimental Artifact #18: Sloppily conducted experiments (which requires no definition).

As mentioned previously, the text has purposefully avoided any detailed discussion of regulatory issues, so if you decide to provide more detail regarding your institutional requirements or your discipline’s standards this would be a good place for that content. One aspect of this topic that many experienced investigators fail to acknowledge is the fact that the sometimes irritating detailed record keeping required by IRBs and funding agencies is designed to increase the quality of the research conducted under their aegis. If there is a planned class experiment, the need for careful attention to detail in every phase of the experiment’s conduct should be stressed. If, as recommended earlier, you or a guest lecturer describes one or more of your (their) personal experiences conducting a favorite experiment, this, too, would be an excellent place to stress the need for the following:

1. Attention to detail,
2. The use of checklists,
3. Pilot studies,
4. Supervision of staff (or helpers for unfunded studies),
5. Training of staff (or helpers),
6. Staff meetings,
7. Using standardized scripts for instructions to participants, and
8. Anticipation of problems, questions that participants might have, and so forth.

And perhaps most importantly of all, all of theses elements need to be documented, because as they used to say in quality assurance: “If it’s not documented, it wasn’t done.”
SAMPLE QUESTIONS

Q1. Which of the following best characterizes a manual of operating procedures (MOP)?
   [A] It contains step-by-step instructions for the conduct of a specific experiment.
   [B] It contains the statistical procedures that will be used to analyze the experimental data.
   [C] It contains the scientific justification for the study hypotheses.
   [D] It contains detailed information on the numbers of participants recruited, consented, and illegible, and refusals to participate (along with reasons given if possible).

Answer: [A] Although definitions may vary, the MOP normally consists of a detailed procedural description of all tasks that need to be performed during the course of the experiment. It does not typically contain statistical procedures or the scientific background of the study. And while recruitment procedures will be specified in the MOP, detailed information on the results of the recruitment process will not be recorded in it. Information of this sort will be stored in secure data files and analyzed as part of the experimental record.

Q2. Which of the following must be revealed to individuals before they agree to participate?
   [A] The hypothesis to be tested
   [B] The rationale for the intervention group
   [C] Risks involved in participation
   [D] Their personal group assignments
   [E] The amount of their time participation will require

Answer: [C] and [E] Research participants must be informed regarding potential risks and the amount of time and effort (as well as the number of assessments) that will be required of them. They need to be informed of the general purpose of the study and their right to withdraw at any time for any reason, but not the specifics of the experimental conditions or the specific scientific purpose.

Q3. What are acceptable functions of a pilot study?
   [A] To test the effectiveness of the intervention
   [B] To test the procedures for implementing the intervention
   [C] To test the feasibility of data collection procedures
   [D] Staff training
   [E] None of these are reasonable expectations for a pilot study given its sample size

Answer: [B] [C] and [D] Pilot studies are too small to test the effectiveness of an intervention. They may provide encouragement that the intervention will be effective, or they may suggest that the intervention needs to be strengthened (or even abandoned), but their primary functions are to (a) test the feasibility of the experimental procedures and (b) train staff (if possible) in the administration of these procedures. If the investigator is the only individual involved in the study, then the pilot study will obviously provide training for him or her (however, someone else should be involved in the collection of potentially reactive outcome data and self-reported outcomes and that person should be trained).

Q4. If a change in the study protocol is indicated, which of the following is or are absolute necessities?
   [A] Notify the IRB.
   [B] Conduct another pilot study.
Conducting an Experiment

1. Truncate the experiment and begin again.

2. None of these, because there are no excuses for changing a study protocol during the course of an experiment

Answer: [A] While it is optimal not to change a study protocol during a study (and certainly if these changes involve major alternations to the intervention, comparison group, or outcome variable then [C] might be necessary), minor protocol changes are not that uncommon (such as adding another questionnaire that might clarify the etiology of an effect). For any of these contingencies, it is probably too late to conduct another pilot study.

Q5. What is the primary purpose of a codebook?

[A] To facilitate data entry
[B] To serve as a reference document for coding decisions made during the course of the study
[C] To ensure treatment fidelity
[D] None of these

Answer: [A] and [B]. A codebook is a necessity for data entry. Also, changes in coding almost always occur in complex studies as unanticipated situations arise.

Q6. Treatment fidelity refers to the process of:

[A] Identifying participants who need to be dropped from the final analysis because of noncompliance
[B] Measuring and documenting participant compliance
[C] Ensuring the proper implementation of the experimental conditions

Answer: [B] and [C] Dropping participants from the final analysis because they did not comply with the treatment protocol is usually no longer an option once they have been randomized because of the growing expectations that intent-to-treat analyses will be conducted. Treatment fidelity is useful in identifying which participants did and did not receive the full benefit of the intervention, which in turn permits subgroup analyses (as long as they are clearly identified as such and not overemphasized) involving participants who complied with those who did not. Treatment fidelity checks during the course of the experiment are also helpful in correcting lapses in staff performance or deviation from the MOP.

Q7. Who bears the ultimate responsibility for supervising the proper conduct of an experiment?

[A] The project director
[B] The IRB or funding agency
[C] The investigator
[D] Both the investigator and the statistician or data analyst (if the two roles are separate)

Answer: [C] The project director may have day-to-day supervisory duties with respect to other staff, but the ultimate responsibility resides with the investigator, and it is the investigator who supervises the project director. [D] may be a de facto answer in some studies in which a statistician is in charge of data entry personnel, but ultimately it is the investigator's responsibility to ensure appropriate and accurate data entry (and he or she should at the very least request data reports involving accuracy checks).

Q8. Which of the following is (are) advantage(s) of regularly scheduled staff meetings?

[A] The identification of unanticipated problems with the implementation of the protocol
Q9. Which of the following are recommended for ensuring accurate data entry?

[A] Double data entry
[B] The periodic computation of descriptive statistics during the course of the entry
[C] Having the data analyst conducting the final analytic procedures promised in the proposal several times during the course of the study
[D] None of these are likely to affect data entry accuracy.

Answer: [A] and [B] Entering the data twice by two different individuals, while time consuming, is the most effective strategy known for ensuring the accuracy of data entry. Periodically running descriptive statistics including the highest and lowest value of each data point is also effective in detecting outliers as well as providing a gross check on accuracy. As mentioned several times, continually running inferential statistics during the course of an experiment is not recommended.

Q10. What steps are important for a principal investigator to take to ensure the appropriate analysis of the experimental data?

[A] Personally analyze the data.
[B] Seek a second opinion about the analysis.
[C] Obtain a writable copy of the results in order to make notes and clarify points.
[D] None of these if the data are analyzed by a statistician

Answer: [C] I believe that [A] is optimal for relatively small-scale experimental studies, but it is not always practical (and it is difficult for non-statisticians to keep up with changes in style in data analysis). There are collaborative situations in which an investigator grows very confident in a statistician’s judgment over the course of a long collaboration. Even then, however, it is important for an investigator to have at least a conceptual understanding of the analyses performed and a record of questions and their clarifications regarding these analyses in the event that (a) the collaboration ends for one reason or another or (b) the data analysis is revisited in later years. (A second opinion need be sought only in the event that the statistician or data analyst’s explanations aren’t fully understood or agreed with.)
This chapter is more informational in nature than practice oriented. The concept of external validity is explained, along with its original definition in the form of the question: “To what populations, settings, treatment variables, and measurement variables can this effect be generalized?”

Many instructors may find the text’s take on the construct a bit too iconoclastic and obviously many scholars and instructors will take issue with it, so please feel free to ignore or correct my opinions on the matter. The Green and Glasgow (2006) insert, much of it comprised of direct quotes, is quite detailed, helpful, and noncontroversial. The article itself might be useful to assign for instructional purposes: http://www.seattleimplementation.org/wp-content/uploads/2011/12/Green-Glasgow-2006.pdf.

Five empirical strategies for assessing and enhancing external validity (or the generalizability of experimental findings) were also presented:

1. **Pragmatic** (aka large simple, scaled-up, or practical) **trials**, which are designed to ascertain if an effect initially demonstrated under controlled conditions can or will be implemented in clinical practice.

2. **Multicenter experiments**, in which the same experimental procedures are implemented at different sites as part of the same experiment.

3. **Meta-analysis**, which is a method by which individual experiments addressing the same (or very similar) intervention (but usually using different populations, conditions, and methodologies) involves the analysis of these experiments as a single group as well as in comparison to one another.

4. **Purposeful replications of experiments by independent investigators**, which are relatively rare (at least as a statement of purpose) in many of the disciplines involving research on humans.

5. **Translational research**, which often involves converting a laboratory-generated finding to a more veridical setting.
**CLASS DISCUSSION SUGGESTIONS**

Once the chapter has been read and this content covered, some discussion by the class on the concept of external validity may be both interesting and helpful. Students should be ready for this discussion since they will have read and discussed several experiments in their own disciplines and most likely have designed some hypothetical experiments of their own.

The topic of external validity overlaps the border between practice and philosophical scientific issues, thus the subject matter of this chapter is a little grayer than most of the topics covered in the book up to this point. The purpose of the brief section entitled “In Praise of Well-Conducted, Laboratory-Type Experiments,” in fact, was included partly to facilitate a discussion to underline the subjective nature of the very importance of external validity. In addition to critically discussing this section, some questions posed to the class (or which might be assigned for *brief* individual or group essays/presentations) might be as follows:

1. Given that Green and Glasgow’s 16 reporting requirements are quite comprehensive, do you think they are realistic? Necessary? Why? (It wouldn't hurt to remind students that these are suggested reporting requirements to help the research consumer assess external validity for their own needs.)
2. What do you think about Lucas and Zelditch’s view of the role of external validity and theory? Does it apply to your area of interest? How?
3. What do you think about the author of their text’s rather negative take on the utility of the external validity concept?

**SAMPLE ITEMS**

**Q1.** Which statement follows from the classic conception of external validity (i.e., addresses issues of generalizability to different populations, settings, treatment variables, and measurement variables)?

[A] If an experimental finding is important enough it will be replicated.

[B] Replication under identical conditions does not address external validity as defined by this definition.

[C] External validity is a multifaceted concept, all of which cannot be addressed in a single experiment.

[D] None of the above is true.

*Answer:* [B] and [C] Since the classic definition involves generalization across participants, settings, measurement techniques, and even interventions, it is difficult to imagine how they could all be addressed in one replication or a single experiment. Obviously, an experiment that cannot be replicated does not possess external validity.

**Q2.** Internal validity is

[A] A necessary condition for external validity

[B] A sufficient condition for external validity

[C] Completely independent of the external validity concept
External Validity

*Answer:* [A] If the original experiment does not produce a defensible inference, this inference cannot be generalized. If the experiment constituted a well-controlled laboratory study, however, it still might not be generalizability to non-laboratory conditions (hence internal validity is not a sufficient condition of external validity).

**Q3.** Which of the following constitute primary, planned differences between an efficacy and a pragmatic trial?

- [A] The experimental purpose
- [B] The experimental setting
- [C] How the intervention is implemented
- [D] None do

*Answer:* [A] [B] and [C] Very liberally translating Zwarenstein, Treweek, Gagnier, et al. (2008), the basic purpose of an efficacy trial is to answer a “can” question while a pragmatic trial address a “does” question. As for settings, an efficacy trial often employs a well-resourced, “ideal” setting as opposed to normal practice. In an efficacy trial, the assurance of treatment fidelity is a primary concern, while many pragmatic or practical trials are designed to ascertain the degree the intervention will be implemented.

**Q4.** Assuming concordant findings, which of the following can provide evidence of the external validity of an experiment?

- [A] Replication of the experiment by independent investigators
- [B] A multicenter trial
- [C] A meta-analysis involving the same variables
- [D] A practical/pragmatic trial
- [E] None do

*Answer:* [A] [B] [C] and [D] None of these strategies guarantee external validity, but all provide some comfort regarding generalizability—replication, because if the original inference is not reliable due to a Type I (or other type of) error, external validity (or generalizability) is irrelevant; multi-center trials, while not exactly independent replications, are at least conducted in different settings; meta-analyses to the extent that the most similar studies to the one in question produce the same inference; supportive practical/pragmatic trials (which are a subset of translational research) perhaps provide the most assurance if their results support the original findings on which they are based.

**Q5.** Which of the following relates to the external validity of an experimental test of a theory?

- [A] The experiment’s relevance to the theory
- [B] Whether or not the theory has been proven to be valid
- [C] The match between the experiment’s intervention and outcome variable and the theory’s constructs

*Answer:* [A] and [C] Theories are never *proven* to be valid, they are seldom even accepted by everyone in a given field, at least until the opposition dies off (to paraphrase James Clerk Maxwell), and by then several new theories have probably arisen. The authors cited in this chapter (Lucas [2003] and Zelditch [2007] who have written on theory testing argue that external validity as defined by Campbell and Stanley and others only applies to theories that specify applicable (or nonapplicable) populations, settings, and/or other relevant experimental conditions.
CHAPTER 12

THREE ADDITIONAL EXPERIMENTAL PARADIGMS

Single-Case, Program Evaluation (Including Natural Experiments), and Quality Improvement Research

The purpose of this chapter was to introduce three unique experimental paradigms (single-case experiments, quality improvement studies, and program evaluation), none of which routinely employ randomized designs, and all of which are characterized by somewhat different philosophical objectives than typical efficacy trials (not the least of which involves less concern with the generality and external validity of findings).

SINGLE-CASE EXPERIMENTS

Single-case experiments, as their name implies, rely on individual participants for their inferences and employ repeated measurements (and often repeated introduction of the same interventions) to replace the control afforded by randomizing multiple participants to multiple conditions. (Ironically, although they are often described as N of 1 studies, they usually employ the results from three or more participants presented separately [usually graphically] rather than analyzed as a group.) Several different strategies for achieving this control were discussed, including the ABAB, ABABAC, staggered-intervention, and changing-criterion designs.

PROGRAM EVALUATION

The second paradigm, program evaluation, is designed to ascertain the “value” of a predeveloped intervention designed for a prespecified population (or institutional setting). Often randomization is problematic, hence the use of quasi-experimental designs is more common than
is the case for conventional efficacy experiments. Also, evaluation researchers tend to have less interest than conventional investigators in testing theories or generalizing their results to other settings or samples.

The regression discontinuity design, though quite specialized, is one of the more robust of these evaluative strategies. Its rationale is not particular intuitive, hence after introducing the logic behind the design, it might be worth assigning students the task of composing a scenario in which the strategy might be employed in their area.

NATURAL EXPERIMENTS

Conceived of here as a subset of evaluation research, natural experiments probably deserve more coverage than I was able to provide in the overview chapter. Their Wikipedia entry (http://en.wikipedia.org/wiki/Natural_experiment) provides a number of examples, as does a book chapter by Remler and Van Ryzin (2010), both of which would be reasonable extra reading assignments (http://www.sagepub.com/upm-data/33935_Chapter13.pdf).

QUALITY IMPROVEMENT RESEARCH

In some ways the final paradigm, quality improvement experimentation, may seem antithetical to the carefully controlled approaches to experimentation advocated in previous chapters—tending as it does to emphasize neither generalizability nor randomization of participants to experimental conditions. It might be interesting to ask students if they see any place for this “attitude” in the conduct of meaningful empirical work in their discipline. They should also be queried about the potential utility of the two strategies or designs (control charting and benchmarking) briefly overviewed in the text.

SAMPLE ITEMS

Q1. What most commonly distinguishes an ABAB single-case design from a randomized crossover design?

[A] Single-case designs normally employ more observations than crossover designs.

[B] Single-case designs normally interpret their results participant by participant rather than by the use of group means (or other summary statistics).

[C] Single-case ABAB designs employ no clear comparisons with which to evaluate the intervention’s effects.

Answer: [A] and [B] The repeated assessments of the outcome before and after the intervention serve as the interventions’ comparisons in an ABAB design. Although there are exceptions, single-case experiments normally employ more assessments than crossover designs.
Q2. How might the following finding be interpreted (see Figure IM.11)?

[A] The outcome variable was inappropriate in the sense that it did not return to baseline.
[B] The intervention effect was of longer duration than expected.
[C] The outcome change observed from the first baseline (A) to the intervention phase was due to an experimental artifact of some sort.
[D] None are plausible explanations.

Answer: [A] [B] and/or [C] could be the case in the absence of any additional information, although hopefully enough pilot work had been done to exclude [A] and [B]. One reason that single-case designs are normally repeated several times is so that “additional information” is available for results such as this. [C] could result from an experimental artifact such as an extraneously occurring external event or a Hawthorne-like effect, although if the same pattern occurred for, say, three participants (who would presumably be run at different times) some experimental artifacts could probably be discounted.

Q3. Would an ABABC design have been preferable here?

[A] Yes
[B] No

Answer: [B] Probably not, since the outcome variable showed no tendency to revert to its original baseline values.

Q4. If this was your experiment and you obtained the same results, what do you think the most responsible course of action would be?

[A] Replicate the experiment.
[B] Write the results up and submit them for publication.
[C] Abandon the entire project.

[D] Conduct more pilot work to ascertain if the results could be explained by problems with the outcome variable or if the intervention was longer lasting than predicted.

[E] Redesign and rerun the study as a randomized controlled experiment.

Answer: [A] [D] and [E] could all be sensible options, depending on the circumstances. [D] is preferable if there is a possibility that the intervention is longer lasting than predicted (obviously a meaningful finding). I would argue that [C] is probably an inappropriate course of action since, if the experiment was important enough to run in the first place, it is probably worth following up. For an efficacy trial involving a large number of participants, [B] would be recommended, but one of the advantages of a single-case study is that the investigator can mount follow-up experiments without a great deal of resource expenditure. (The investigators should probably mention or include the first experiment in their final paper, however, if they choose option [D].)

Q5. The staggered-intervention design is preferable to an ABAB design when

[A] The intervention is of relatively brief duration.

[B] The investigator does not feel it is ethical to withdraw the intervention if it is successful.

[C] The outcome variable is not expected to return to baseline.

[D] None of these apply.

Answer: [B] and [C] If the outcome does not return promptly to baseline, which could be a characteristic of the construct being measured or the intervention's lasting power, the ABAB or similar design would be contraindicated. The ABAB design would also be contraindicated if the investigator felt that the benefits of the intervention were too important to withdraw from the participant(s).

Q6. What is the most common distinguishing characteristic of program evaluation (as opposed to conventional efficacy experiments)?

[A] The types of designs employed

[B] The number of participants used

[C] The types of outcomes employed

[D] The process of creating the intervention

Answer: [D], because program evaluations normally use interventions that are designed and often implemented by someone other than evaluator. [A] is an understandable answer as randomized designs are rare in evaluation. Evaluation studies also often also use more participants than the average efficacy trial, but [D] constitutes the most distinguishing characteristic of evaluation studies.

Q7. Which of the following is or are not considered a function of the design program evaluation process?

[A] Engage stakeholders in the evaluation.

[B] Design and focus the evaluation.

[C] Collect data (or gather “credible” evidence).

[D] Take steps to ensure use of the evaluation.

[E] All are reasonable functions.

Answer: [E], at least from the perspective of the Centers for Disease Control and Prevention (a major funder of the genre).
Q8. How is the regression discontinuity design distinguished from a nonrandomized pretest posttest control group design?
[A] The regression discontinuity design does not use a pretest and posttest.
[B] The regression discontinuity design assumes that the two groups being compared are nonequivalent.
[C] The nonrandomized pretest-posttest control group design assesses differences between groups, while the regression discontinuity design assess differences in relationships.
[D] None of these apply.

Answer: [B] The regression discontinuity design is employed when one “group” of individuals receives an intervention because they possess higher (or lower) scores on a selection variable (which is comparable to a pretest) than the other “group,” hence the groups are by definition nonequivalent on that variable/pretest. This particular design is unique, however, because it is applicable only when a specific cut-point can be defined on the selection variable where everyone on one side of the cut-point receives the intervention while everyone on the other side is denied the intervention. Both designs assess between-group differences, but the regression discontinuity design assesses whether or not there is a break (or disconnect) in posttest scores (based on the pretest-posttest regression line) at the pretest cut-point.

Q9. What type of outcome would be preferable for a six sigma control chart?
[A] One that occurred continuously over time
[B] One that was not reactive
[C] One that was routinely collected
[D] One that was not subjective in nature

Answer: [A–D] Control charts require a constant flow of data which are preferably (a) routinely collected (or which should be routinely collected) as part of the delivery of a service or the process of production and (b) objective (and certainly not personal responses or self-reports).

Q10. Benchmarking is:
[A] A commonly used strategy in single case research designs
[B] A commonly used strategy in quality improvement studies
[C] A strategy for identifying promising interventions for implementation in quality improvement studies
[D] None of these

Answer: [B] and [C] Most often associated with total quality management and quality improvement, the strategy involves identifying effective strategies for improving the quality of services (or care) delivered in exemplary settings and translating them to settings in need of improvement. (Benchmarking is also used in evaluation, often to identify realistic outcome goals.)
CHAPTER 13

EXPERIMENTAL BIAS

Two types of bias were discussed:

1. *Publication bias*, which is a propensity for experiments with statistically significant findings to have a greater likelihood of being published, and
2. *Investigator bias*, which is a conscious or unconscious misrepresentation of experimental results ranging from prosecutable fraud to fabrication of data to not reporting certain negative aspects of a study.

The first priority in covering these topics is to ensure that students are aware of what constitutes scientific misconduct. The second should be to alert them of the presence of experimental bias and the danger it poses to their own work.

ACTIVITIES

*Student reports of historical examples of misconduct:* This chapter has barely scratched the surface of the fraud and misconduct cases that have been reported and analyzed over the past several decades. A reasonable assignment might entail students selecting examples of fraud or misconduct, writing a few paragraphs on them, and describing the cases to the class. Most are available and reasonably well described on *Wikipedia* (see the link below), as well as books such as the following:


Other classic examples mentioned in the text include the Darsee scandal, William T. Summerlin, Cyril Burt, and the cold fusion fiasco. Egregious violations of participant rights are another source of scientific misconduct, such as the Tuskegee syphilis study, or less
gruesome but controversial studies such as Milgram’s obedience studies. No shortage of examples exists, such as the following exhaustive treatment in Wikipedia: http://en.wikipedia.org/wiki/Unethical_human_experimentation_in_the_United_States.

Rating articles on their methodological quality: As mentioned in the text, one group of investigators (Moyer, Jadad, Nichol, et al., 1995) located 25 instruments designed to rate the quality of research studies, thus there is no lack of rating instruments available for students. A reasonable design activity is to assign students the task of rating a small sample of experiments from their own discipline using one of the easier rating instruments available, such as the Jadad scale. Alternately, they could make up their own scales or add items to an existing one. This is both good practice in applying design principles to the published literature and a potentially publishable research activity in its own right if done systematically as part of a class project.

A caveat, however, is in order since many rating instruments are discipline specific. (“Discipline specific” refers not only to meta-categories such as medicine or education but also to specific types of interventions and outcome variables.) The Jadad scale presented in the text, for example, relies heavily on the use of blinding, which is rarely used in some disciplines for which credible placebo groups are difficult to formulate. (However, just because blinding is difficult doesn’t vitiate its importance.) The same is true of the Bausell items (sample size over 50/group, randomization, placebo control group, attrition less than 25%, and high-impact journals) as well as the checklist based on the CONSORT statement, presented next. Therefore, if the class chooses an area in which a published design rating scale is not available a reasonable group activity might be to construct such a scale and critique it, item by item, in class prior to utilizing it with a group of homogeneous experiments.

The 2010 CONSORT (Consolidated Standards of Reporting Trials) Statement and Checklist: Although I have discussed scientific acculturation issues at some length, I have purposefully avoided the nuts and bolts of scientific publishing except for those issues related to transparency, honesty, and scientists’ obligation to publish their work. Should you wish to pursue the ins and outs of publishing in greater detail, a good place to start might be the research standards available through your own professional association (e.g., APA, AERA, ASA, and so forth).

The CONSORT Statement

I personally find the CONSORT (Consolidated Standards of Reporting Trials) Statement the most methodologically comprehensive of those standards with which I am familiar, and I especially recommend their checklist based on the full CONSORT Statement—even though some of its items are discipline specific. The document was developed by a consortium of investigators, biostatisticians, and medical journal editors to improve the quality of reporting (and conduct) of medical trials. The standards are periodically updated, and addenda related to specific designs have been recently added (as cited throughout the text). A checklist based on the full document comprises good experimental practice in medical trials but is applicable to all experiments involving human participants. It has been adopted by the highest impact journals in medicine, and in my opinion all investigators in all fields involved in experimentation with human participants should at least be aware of it.

Both the statement and the checklist are organized around the constituent components of a journal article, with recommendations concerning what should
be included in the introduction, methods, methods, results, and discussion sections as well as the title and abstract to increase transparency and acceptable experimental practice. Naturally, journals in various disciplines will have different structural and stylistic conventions, but the basic advice is all excellent. The entire document is freely available for downloading at http://www.consort-statement.org/ and can be reproduced (including the checklist that follows) as long as appropriate attribution is provided.

<table>
<thead>
<tr>
<th>Section/Topic</th>
<th>Item No</th>
<th>Checklist Item</th>
<th>Reported on page No</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title and Abstract</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td></td>
<td>Identification as a randomized trial in the title</td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td></td>
<td>Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)</td>
<td></td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background and objectives</td>
<td>2a</td>
<td>Scientific background and explanation of rationale</td>
<td></td>
</tr>
<tr>
<td>2b</td>
<td></td>
<td>Specific objectives or hypotheses</td>
<td></td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial design</td>
<td>3a</td>
<td>Description of trial design (such as parallel, factorial), including allocation Ratio</td>
<td></td>
</tr>
<tr>
<td>3b</td>
<td></td>
<td>Important changes to methods after trial commencement (such as eligibility criteria), with reasons</td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>4a</td>
<td>Eligibility criteria for participants</td>
<td></td>
</tr>
<tr>
<td>4b</td>
<td></td>
<td>Settings and locations where the data were collected</td>
<td></td>
</tr>
<tr>
<td>Interventions</td>
<td>5</td>
<td>The interventions for each group with sufficient details to allow replication, including how and when they were actually administered</td>
<td></td>
</tr>
<tr>
<td>Outcomes</td>
<td>6a</td>
<td>Completely defined prespecified primary and secondary outcome measures, including how and when they were assessed</td>
<td></td>
</tr>
<tr>
<td>6b</td>
<td></td>
<td>Any changes to trial outcomes after the trial commenced, with reasons</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
### Sample size
- **7a**: How sample size was determined
- **7b**: When applicable, explanation of any interim analyses and stopping Guidelines

### Randomization
- **Sequence generation**
  - **8a**: Method used to generate the random allocation sequence
  - **8b**: Type of randomization; details of any restriction (such as blocking and block size)
- **Allocation concealment mechanism**
  - **9**: Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned

### Implementation
- **10**: Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions

### Blinding
- **11a**: If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how
- **11b**: If relevant, description of the similarity of interventions

### Statistical methods
- **12a**: Statistical methods used to compare groups for primary and secondary Outcomes
- **12b**: Methods for additional analyses, such as subgroup analyses and adjusted analyses

### Results
- **13a**: For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome
- **13b**: For each group, losses and exclusions after randomization, together with reasons
- **Recruitment**
  - **14a**: Dates defining the periods of recruitment and follow-up
  - **14b**: Why the trial ended or was stopped
- **Baseline data**
  - **15**: A table showing baseline demographic and clinical characteristics for each group
Numbers analyzed | 16 | For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups

Outcomes and estimation | 17a | For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)

| 17b | For binary outcomes, presentation of both absolute and relative effect sizes is recommended

Ancillary analyses | 18 | Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing prespecified from exploratory

Harms | 19 | All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)

Discussion

Limitations | 20 | Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses

Generalizability | 21 | Generalizability (external validity, applicability) of the trial findings

Interpretation | 22 | Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence

Other Information

Registration | 23 | Registration number and name of trial registry

Protocol | 24 | Where the full trial protocol can be accessed, if available

Funding | 25 | Sources of funding and other support (such as supply of drugs), role of funders

* Note that a few of the items are discipline specific, such as Item 23, which refers to the registration number and name of a trial registry.

**Compiling a list of non-sanctioned scientific behaviors:** Have students compile a list of questionable scientific behaviors, classify them in some way (e.g., fraud vs. misconduct, or censurable vs. non-censurable) or rate them on a scale from, say, 1 (not really that bad) to 10 (absolutely egregious). Here are 16 behaviors reported by Martinson, Anderson, and de Vries (2005) based on a survey of 3,247 NIH funded scientists. The wording of the items is identical to that presented in Table 1 of the *Nature* article (Martinson, B.C., Anderson, M.S., & de Vries, R. [2005]. Scientists behaving badly. *Nature, 435*, 737–8).
1. Falsifying or “cooking” research data
2. Ignoring major aspects of human-subject requirements
3. Not properly disclosing involvement in firms whose products are based on one’s own research
4. Relationships with students, research subjects, or clients that may be interpreted as questionable
5. Using another’s ideas without obtaining permission or giving due credit
6. Unauthorized use of confidential information in connection with one’s own research
7. Failing to present data that contradict one’s own previous research
8. Circumventing certain minor aspects of human-subject requirements
9. Overlooking others’ use of flawed data or questionable interpretation of data
10. Changing the design, methodology, or results of a study in response to pressure from a funding source
11. Publishing the same data or results in two or more publications
12. Inappropriately assigning authorship credit
13. Withholding details of methodology or results in papers or proposals
14. Using inadequate or inappropriate research designs
15. Dropping observations or data from analyses based on the belief they were inaccurate
16. Inadequate record keeping related to research projects

Note that some of these involve outright fraud (e.g., falsifying data), some are ethical (circumventing certain minor aspects of human-subject requirements), some are publishing abuses (inappropriately assigning authorship credit), and some are plagiaristic (using another’s ideas without obtaining permission or giving due credit). Naturally, you can delimit or add to the list anyway you please or assign groups different categories for the same purpose. Other forms of scientific misconduct (some of which appear in the text) include conducting research as seeding trials, using one’s scientific credentials to further political causes or simply for profit, putting others on one’s publications in return for reciprocal authorships, ad nauseam.

SAMPLE ITEMS

Q1. Publication bias affects the scientific record by:
   [A] Increasing the proportion of false positive results in the experimental record
   [B] Increasing the proportion of false negative results in the experimental record
   [C] Decreasing the amount of information in the experimental record on interventions that do not work and that fail to confirm theories
   [D] Increasing the prevalence of fraudulent results in the experimental record

Answer: [A] and [C]. Since publication bias refers to a preference for publishing statistically significant results to the exclusion of nonsignificant ones, the end result is the appearance that more interventions are effective in a discipline than ineffective.

Q2. Publication bias is an example of:
   [A] Scientific fraud
   [B] Scientific misconduct
   [C] A flaw in the peer review process
**Answer:** [C] Some professionals (and research ethicists) consider it unethical for a scientist not to at least attempt to publish a study that did not support the investigator’s hypothesis, but few would go so far as to call it fraudulent or misconduct. Instead, it is a correctable flaw in the peer review process. It is also a misconception on the part of investigators since perseverance will usually result in a well-designed, negative experiment being published somewhere.

**Q3.** What are some other possible implications of publication bias?

**Answer:** This is an opinion question, but in addition to the ones just presented is the possibility that it discourages investigators from being adventurous with respect to exploring new avenues of research.

**Q4.** Which of the following constitute(s) scientific and/or ethical objection(s) to seeding trials?

- [A] They are based on fraudulent data.
- [B] They are often based on substandard methodologies.
- [C] They possess an inherent conflict of interest.
- [D] They normally do not adequately test a legitimate hypothesis.
- [E] They take advantage of participants’ altruistic motives.

**Answer:** [B] [C] [D] and [E]. [A] (the presence of fraudulent data) occurs, but there have been no studies of which I am aware that document the prevalence of data fabrication in this genre of research. Seeding trials often do not randomly assign participants or use appropriate control/comparison groups. Since their underlying purpose is to sell something in order to profit from it, they also fit all sensible conflict of interest definitions. And since they obviously don’t announce these profit motives, they violate the public’s often altruistic motivations for participating in the studies.

**Q5.** What design characteristics have been found to be significantly related to the size of experimental effects in medical research?

- [A] Placebo vs. non-placebo control groups
- [B] Randomized vs. nonrandomized designs
- [C] Experimental attrition
- [D] None have

**Answer:** [A] [B] and [C] Bausell (2009) found [A] and [C] to be related (only randomized trials were included in his analysis) to the obtainment of statistical significance in trials evaluating alternative medical treatments. Chalmers, Celano, Sacks, and Smith (1983) found [A] and [B] related to reported mortality rates over 35 years ago in reviewing treatments for acute myocardial infarction. Other examples are cited in the text.

**Q6.** What design characteristics have been found to be significantly related to the size of experimental effects in behavioral and educational experiments?

- [A] Placebo vs. non-placebo control groups
- [B] Single-group vs. multigroup designs
- [C] Randomized vs. nonrandomized designs
- [D] None have

**Answer:** [B] The most definitive analysis of behavioral and educational experiments of which I am aware is Lipsey and Wilson’s *American Psychologist* article (1993). There they found that single-group studies reported higher ESs than experiments using control groups, but no such effect for randomized vs. nonrandomized multigroup designs. (They also found that dissertations...
InTroducTIon T o THE EXPERIMEnT AL ProcESS

reported smaller ESs than unpublished experiments, which is normally attributed to publication bias.)

Q7. Long-term methods for evaluating the credibility of an experiment include:
   [A] Citation counts
   [B] Independent replications of experiments
   [C] Consideration of the scientific plausibility of theories on which experiments are based
   [D] None of these

Answer: [B] and [C] Independent replications are somewhat rare except for high-stakes research such as the cold fusion example. An experimental effect with no plausible theoretical rationale (e.g., homeopathy) should always be viewed with extreme skepticism.

Q8. The only proven/foolproof method(s) of detecting fraudulent experimental practices is (are):
   [A] Using validated, discipline-specific experimental design rating systems
   [B] Independent replication
   [C] Excessive hype of the part of the authors
   [D] None of these

Answer: [B] Although not perfect, the failure of multiple transparent independent replications is one of the best approaches we have. Others are vigilance (such as occurred in the Cyril Burt case) and whistle-blowing (which can be more damaging to the whistle-blower’s career than the perpetrator of the fraud).
A CLOSING WORD

I hope that you have found some of the suggested activities and sample items useful, and I appreciate your adopting the text. Most of all, I wish you continued success in your own research activities and especially continued success in training our next generation of experimental scientists.