

Teaching Experimental Design

Robert G. EASTERLING

After a career as a consulting statistician, I decided to teach. This article describes my evolving views as I developed and taught introductory university courses on experimental design, organized around and stimulated by three different texts and three different universities. Primarily, I found it to be essential, particularly at the course's beginning, to embed textbook examples in credible scientific or business contexts in order to try to convince students of the value of statistical experimental design and analysis in their subsequent careers. In contrast, uninteresting, even nonsensical, undeveloped examples that serve only as formula drill give students the opposite impression. My purpose in this article is to present selected expanded textbook examples and to use these illustrations to examine fundamental issues in experimental design such as: the importance of subject-matter, the choice of experimental units, the nature and purpose of blocking, and the contrast between random sampling and random assignment of treatments. These examples expose fundamental issues in how our profession functions in a collaborative environment and how we prepare the next generation of statisticians and statistically savvy professionals. My hope is that the illustrations provided will be directly useful to beginning instructors and provocative to the experienced.

KEY WORDS: Blocking; Experimental units; Randomization; Replication; Students; Statistical thinking.

1. INTRODUCTION

So I decided to teach. After 34 years as a consulting statistician at Sandia National Laboratories I decided my next career would be as an occasional roving visiting professor. My motivation was a bit grandiose: to teach what I've learned about the practice of statistics to the next generation. Besides, I've met enough people who say that statistics was the worst/hardest/most useless/most boring course they've ever taken that it became a challenge to see if I could counter this impression.

In Spring 2000, while still employed by Sandia, I taught a first course in experimental design at the University of New Mexico (UNM). This course was for seniors and graduate students in statistics and other fields. I used the classic Box, Hunter, and Hunter (1978) text because I felt all budding statisticians and experimenters ought to have this book on their desks. The UNM

Until July 2001, Robert G. Easterling was employed by Sandia National Laboratories in various consulting and management positions. Since then he has been an itinerant visiting professor. He thanks the Universities of New Mexico, Michigan, and Auckland for the opportunity to teach and to sample life in their communities. Thanks also to the editor, associate editor, and reviewers for their comments that helped improve this article. Address: 51 Avenida del Sol, Cedar Crest, NM 87008 (E-mail: rgeaste@comcast.net).

experience was positive enough that I decided to retire from Sandia and take my show on the road.

My first stop was the University of Michigan, Fall 2001, where I taught a similar level of experimental design course, this time using Montgomery's (2001) text. I changed and tried to improve on how I had presented things at UNM and plugged in some of Montgomery's examples. (One thing I decided early on was that if the university asked students to buy a text, then I should help them get the most possible value out of it. Originally I had planned to use Sandia real-world case studies.) The two texts are structured similarly so course organization was basically unchanged. Class size was 50, rather than the 14 I had at UNM.

Still feeling positive about this odyssey, I next taught an introductory course in experimental design at the University of Auckland, Spring 2003 (in the northern hemisphere). At the University of Auckland the course is based on internally produced Lecture Notes (Scott and Triggs 2003), the experimental design half of which was written by the course developer and regular instructor, Chris Triggs. The student body was now 105, more heavily weighted toward third-year students and a smaller percentage of graduate students than my previous courses, and the course was shortened—I had 24 hours of lecture versus 36 at the previous stops. The course Lecture Notes are organized in a somewhat novel way that I found I liked (discussed below) and I rearranged and reformulated my lectures to follow Triggs' Notes.

At all three universities students had taken the mandatory Stat 101 class and, in some cases, a second statistics class in regression or sampling. Much has been written about our Stat 101 classes and much innovative work has been done in this area. However, the negative popular impressions of statistics cited above can often be traced to unenlightened Stat 101 courses. One instance from around 20 years ago stands out in my mind as illustrative of this problem.

In browsing through a problem section of a new text at the time I came across this problem (paraphrased):

A shoe-store owner records the shoe sizes for the last 20 women's-shoes purchases in his store. Data are given. The student/victim is told: Test the hypothesis that the population median shoe size is 7.5.

What!? Is there any sensible reason why a shoe store owner would do this? Would you want to work with or for someone who thinks this is what statisticians do? I'm reminded of a remark by the late and great W. E. Deming: When told that the biggest problem Stat 101 students have is recognizing whether a hypothesis test is one-sided or two-sided, he remarked, "Maybe they are trying to think!"

The sole purpose of exercises such as this (I call them self-inflicted wounds on our profession) is formula drill, not statistical thinking, not rational thinking. It's hard to teach thinking, or write about it, or create problems that exercise it. It's much easier to teach formula or software-script plug-ins. ("Monkey

see; monkey do” is how one professor characterized it to me. How demeaning of our students! This is no way to gain their respect.)

With this background, I have increasingly tried in my (limited) teaching to embed textbook examples into real-world-like contexts. There can be economic, environmental, and ethical issues involved in deceptively simple, even made-up, experiments pertaining to, for example, what fertilizer to use on tomatoes or what shoe leather to use in boys’ shoes. I try to raise the students’ consciousness in this regard in the first lecture and throughout the course. My goal is to prepare at least some of them for careers in which well-aimed, well-run, well-analyzed experiments are used (with enthusiasm) to generate useful information that leads to rational actions and consequences.

This is not a novel thought, of course. I’m sure many instructors successfully add meaningful and stimulating context to textbook examples, but I thought it might be directly useful to provide beginning instructors with some illustrations from my recent experience. Also, experienced teachers can benefit from a different perspective on familiar examples and issues. Further, issues that arise in an expanded treatment of deceptively simple examples are issues that all users of statistical methods need to ponder occasionally. Some of these are fundamental issues pertaining to how statisticians function in a collaborative endeavor and how we prepare the next generation of statisticians and their professional colleagues to do likewise. These issues include:

- the importance of the contexts (business, ethical, scientific, financial, . . .) in which experiments are conceived, conducted, and analyzed
- the choice of experimental units
- the nature and roles of blocking
- the nature and extent of replication
- the role of randomization (and the difference between randomization and random sampling)
- the difference between structured observational studies and experiments
- the statistical and subject-matter bases for inferences drawn from the experimental results

My teaching experience has led to changes in how I will teach the subject and in how I will approach experimental design in

practice. I don’t claim to have found a new or better way. My objective is simply to convey how my experience has shaped my thoughts about experimental design. As I have been stimulated by the courses I have taught, I hope readers will be stimulated to refresh or strengthen their thinking on the fundamentals of statistical experimental design and how they are taught.

In the following sections I deal primarily with small, two-treatment experiments from the texts I have used. Although I find some examples better suited to my purposes than others, it is not my intent to evaluate or compare texts. Other users of these texts may have different needs, impressions, and experiences. Authors work under a huge set of constraints that I only dimly imagine. There is a constant tension between depth and breadth of coverage that has to be resolved and I am daunted by and appreciative of the huge amount of work that goes into a statistical text.

2. EXPANDED EXAMPLES FROM BOX, HUNTER, AND HUNTER (1978)

2.1 Example A. Tomato Fertilizer

Experimental design texts generally start with two-treatment experiments for simplicity and because they provide a link with students’ prior exposure to the two-sample paired and unpaired Normal-theory analyses that they’ve done in Stat 101. The Box, Hunter, and Hunter (1978) (hereafter abbreviated BHH) example is a gardener’s tomato-growing experiment, the objective of which is to “discover whether a change in the fertilizer mixture . . . would result in improved yield.” Eleven plants are set out in a single row and a random assignment of fertilizer A to five plants and fertilizer B, which is the change being considered, to six plants is made (right off I wonder if there was a sixth plant assigned to A and it got stepped on or infected and didn’t survive the experiment. I also raise the question of why 11 plants, not 50, or 100, or . . ., only to talk conceptually about the difficult problem of sizing an experiment that will be addressed later.) BHH use this example to motivate and illustrate a randomization test to evaluate the observed difference between fertilizers.

In setting up this problem, I add some points on protocol: the plants are adequately isolated so that the fertilizer on one plant does not bleed over onto its neighbors. Also, attention must be given to watering, weeding, and insect treatment during this experiment to make sure these variables do not create a bias. A point I hit repeatedly is that there is more to an experiment than just a table of data awaiting your analysis. You have to know about the care and feeding of the experimental units that provided the data.

One feature of this example is that the tomato yields are identified by row position of each plant. The row positions are given to facilitate the discussion of randomizing the fertilizer assignment via shuffling and dealing red and black cards, but I include row position in the analysis. An early lesson that I try to teach is that data-snooping and asking questions can turn up ancillary information that is an important part of extracting and understanding the message in the data. For this experiment I plot the tomato yields, by fertilizer, versus position (Figure 1). (Analysis One, I tell my students, is “Plot the data.”) Two features leap out: a decided declining trend with respect to position and one visibly outlying point (which would not be visibly outlying without as-

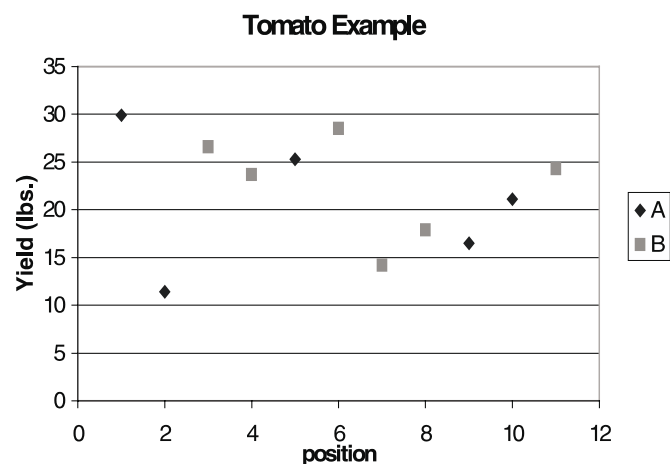


Figure 1. Plant yield versus position, by fertilizer.

sociating yields with position—another point to expand upon). It is fairly obvious from the plot that there is no consistent advantage for either fertilizer. It's more important where you plant your tomatoes than which fertilizer you use! The analysis could stop here. Of course, however, we need to commit some additional statistical analysis to substantiate what we can see. That's what statisticians do.

BHH use this example to illustrate the principle of randomization. The random assignment of treatments to experimental units validates the randomization test of the mean difference in yields for the two fertilizers. Further, the good agreement here with the Normal-theory t test motivates regarding the conventional Normal-theory t test as an (often) adequate approximation to the randomization test. Though not part of the BHH discussion of this example, the example also shows how randomization can provide protection against unrecognized sources of variability. I make the point that if a convenient allocation of treatments had been done—say A at one end of the row, B at the other—misleading results would have occurred.

The conclusion from this small experiment is that there is no (appreciable) evidence of a difference in average tomato yield due to fertilizer. Is that all there is? No. I discuss the actions that might follow this outcome. A negative result can have positive implications. If I'm trying to decide what fertilizer to use in next year's garden, I can make the choice based on other considerations such as cost, environmental impact, and tomato taste. On the other hand, if I'm a major commercial grower of tomatoes, this small experiment has not provided nearly enough information to make a choice when the wrong choice could have major financial effects. The indefiniteness of the results is reflected by the confidence interval on the underlying expected difference in yields for the two fertilizers. It spans zero and the possible range of average yield (lbs./plant) is broad enough that if a commercial grower were to use this experiment to choose a fertilizer, there is a substantial risk of choosing a loser. I return to this matter in a later lecture on experiment size (replication) and use a made-up economic analysis to help drive the size of the experiment.

2.2 Example B. Boys' Shoes

The BHH example for introducing the principle of blocking concerns the comparison of two shoe sole materials used in boys' shoes: A, the standard material; and B, a candidate cheaper substitute. Two designs might be contemplated:

1. For some group of boys, randomly assign half of them shoes with material A; the other half get shoes with material B. After some period of wear, measure the amount of wear on the shoes and compare the data for the two materials.
2. Have each boy in a selected group of boys wear one shoe of each material, randomly assigned to left and right feet.

The choice of designs is not discussed in BHH. Their starting point is that experiment 2 has been done. The point they make is that if the analysis is done via the unpaired t test used in the tomato experiment, that would be wrong. The correct analysis should reflect the pairing. (I prefer not to compare the incorrect unpaired analysis to the correct paired analysis in cases such as these. To do so creates the impression in some students that the

choice of analysis is at the discretion of the analyst, to get the best results, rather than that the analysis is driven by the conduct of the experiment.) BHH do a randomization test based on the paired differences and a paired t test to illustrate again the value of randomization and the good approximation one can get to the randomization test by using the Normal theory t test on paired differences.

I elaborate on this example by starting with the above design issue. This example provides an excellent opportunity to talk about the (generally overlooked issue of) choice of experimental units (eu's): Will it be a boy or a foot? and to talk about the benefits that accrue when experimental units can be logically grouped, or blocked, in such a way that treatments can be randomly assigned to eu's within a block, thus permitting treatment mean differences to be compared to within-block error variability. A more precise comparison should be able to be made compared to that which would be made if treatments were randomly assigned to the same group of experimental units without regard to the blocking variable. As noted, though, that's not the crux of the matter here. The alternative design is not a completely random assignment of materials to individual feet ignoring the boys they belong to. This example deals more with the need to be insightful and clever in defining experimental units (the entity that receives an independent application of a treatment). Experiments such as Fisher's (1935a,b) agricultural experiments are better examples of blocking out sources of variation that would otherwise diminish the precision with which treatments can be compared. One could return to the tomato example and suppose that now that the fertility trend in the experimenter's garden has been recognized, a subsequent experiment could be done in which adjacent pairs of plots could be blocked and fertilizers assigned at random to the two eu's in these blocks. BHH, on the general principle that physically adjacent eu's tend to be more similar than widely separated eu's, indeed suggest this design after their discussion of the boys' shoes paired experiment. It is interesting that the data suggest this blocking. But, you have to go beyond the text to discover this pattern.

The result of the analysis of the shoe data is that there is a (statistically) significant difference between the materials: the cheaper material does not wear as well as the currently used material. It is clear from a display of the data that if shoe materials had been assigned to boys, rather than feet within boys, that a significant difference would not have been detected. That's the statistical point about the randomized block design on which most authors focus.

But, I hammer on another point pertaining to this experiment: Life does not end with a significance test. There is a decision to be made about what material the company should use. The difference in wear is quite small: an average difference of 0.4% wear compared to the average wear in this experiment of about 10%. This can lead to the usual discussion of practical versus statistical significance, but I go further. We might extrapolate the results of this experiment to say that the average life of a shoe (defined for the sake of discussion as 50% wear) will decrease by about 2% if we switch to the cheaper material. Surely customers won't notice this minuscule loss in quality. Their shoes might wear out in 358 days instead of 365. But, suppose your company's slogan is Nothing but the Best! Are you going to recommend the cheaper material to your manage-

ment (and maybe hope for some recognition or reward for the cost savings resulting from your cleverly designed experiment)? Or are you going to argue that the company should avoid this first step on the slippery slope that leads to corner-cutting, poor quality, loss of reputation, loss of sales, then bankruptcy? This is a somewhat overblown ethical dilemma, but I tell students that ethical issues often impinge on experimentation. Sponsors, manufacturers, advocates, thesis advisors, government agencies, and other interested parties have agendas and have strong hopes that experiments will be designed and run with results that support those agendas. I'm not being sinister, just realistic. In fact you want people to care about the experiments you help design and analyze. (Some of my biggest career disappointments have come when well-designed, insightfully analyzed experiments have been shrugged off: "Yeah, we knew that.") The statistician on the job has to be able to detect these influences and have the strength of character to ensure that honest experiments, honestly analyzed, are carried out. For example, the choice of factors and factor levels, often treated as givens in texts that focus on the analysis of an experiment's data, clearly can make or break or bias an experiment, regardless of design efficiency or analysis proficiency. (This shoe-leather example and its ethical implications, along with the shoe-store hypothesis-testing exercise above, put me in mind of the provocative article by Irwin Bross (1974), "The Role of the Statistician: Scientist or Shoe Clerk." I recommend it.)

I also use the shoe experiment to contrast randomization and random sampling. It is the random assignment of shoe materials to left and right feet that underlies the validity of the conclusion that the observed difference is quite unlikely to be a chance outcome. There is no reason to assume (pretend) or require that the boys who participated in the experiment were a random sample from any well-defined frame, such as all 12-year-old boys in Ann Arbor public schools, or from some ill-defined population such as all potential customers of the company's shoes. The participating boys were probably what Deming called a "judgment sample," chosen by the experimenter to be representative of a range of boys—large/small, active/inactive, . . . Skateboarders were likely to be excluded because of their unequal shoe wear. Convenience and availability can also be major considerations. The statistical inference provided by the random assignment of shoe materials is that *among these boys* the average difference in shoe-wear is larger than can be plausibly due to the chance treatment assignment. Any inference that the wear difference seen in the experiment applies to some broader population of shoe wearers must be based on the subject-matter informed judgment that the boys in the experiment adequately represented, or even spanned, the sort of customer use and abuse that these shoes might see.

This comparison of randomization and random-sampling leads me into a brief discussion of Deming's (1975) distinction between analytical and enumerative studies. (This discussion was particularly apt in Auckland where the students were simultaneously taking parallel tracks in sampling and experimental design.) I think we, as a profession, seriously overwork the population/sample (enumerative) concept as the fundamental paradigm of statistics. (The shoe-store owner example in the first section is a case in point. What "population" could that example

possibly refer to?) Designed experiments (analytic) seldom fit this mode and it hurts our credibility to ask students or colleagues to pretend experimental data were obtained by random sampling from some vague hypothetical population. Further, I would argue that progress—scientific, technical, societal—is more likely to result from experimentation than from population sampling, so it is particularly important that we have a firm foundation for the statistical and subject-matter underpinnings of the resulting inferences.

It is fortuitous in this shoe example, of course, that the number of treatments is equal to the number of experimental units in a block that has a natural size. I later introduce the Balanced Incomplete Block Design with a little drama: You're the company statistician who came up with the blocked experiment to compare the two shoe sole materials most efficiently. Now, the head of Shoe Research comes to you and says we now want to compare four upper-shoe materials. "Let's see you put four shoes on two feet!" You come up with the BIBD, show how efficiently it works, and reap company fame and fortune for your cleverness.

3. DISCUSSION

Box, Hunter, and Hunter (1978) devoted a chapter to these two examples and used them to make important points about randomization and blocking. As just described, I put even more real-world weight on the shoulders of these examples. They can take it. This elaboration has evolved as I have now taught these examples three times. At U Auckland I used these examples in the first two lectures to try to get across, not only the fundamental design principles of replication, randomization, and blocking, but also the scientific and business contexts in which experiments are conducted, analyzed, and acted upon. As a by-product, these lectures include a review of the supposedly familiar paired and unpaired *t* test analyses that the students have learned in Stat 101, but I try not to let formula-review dominate these introductory lectures.

As noted earlier, at University of Michigan I taught using Montgomery (2001) as the text. This was my choice because in talking to university acquaintances I had heard that students found BHH to be difficult and that Montgomery (2001) was a currently popular choice. I also knew that using it would usefully broaden my experience in teaching experimental design. A quick review indicated that Montgomery's text has much the same structure as BHH but, of course, the examples are different. So, I adopted it. In teaching the course, however, for various reasons, I found I needed to revert to the BHH examples to convey the messages discussed here. Some discussion of the reasons for this decision may be instructive. As mentioned earlier, my intent in this article is not to critique or compare these two texts, but only to show how they have contributed to my thinking about teaching experimental design.

4. EXPANDED EXAMPLES FROM MONTGOMERY (2001)

4.1 Example C. Cement Mortar

Montgomery's (2001) introductory example of a two-treatment experiment, used to illustrate the conventional two-sample unpaired *t* test, is a comparison of two cement mortars: unmodified and modified. The reader is told: "The experimenter

has collected 10 observations on strength for the modified formulation and another 10 observations for the unmodified formulation.” This succinct description, however, does not contain enough information to justify the subsequent two-sample t test analysis. Was one batch of cement with each mortar prepared, then 10 specimens drawn (at random?) from each batch and measured? If so, the experimental unit is a batch and with only one eu per treatment, no matter how many specimens are measured, we have no measure of experimental error against which to compare the two mortars. On the other hand, if 10 separately prepared samples of cement were prepared with each mortar, with mortar type being randomly assigned to the cement mix, then we’ve got an experiment for which the usual comparison of treatment means is valid. So, when I taught out of Montgomery I used this example to make the point again: You can’t tell the nature of an experiment from the data layout on a piece of paper, from the mere fact that we have 10 “observations” of each method. Experimental design matters! Either this mortar experiment is virtually worthless (no replication) or it is very definitive (the two sets of data are completely separated). (The modified mortar is inferior to the unmodified, so I ask, “Whose smart idea was it to modify the mortar?”) Because it is not clear that this set of data should be analyzed as a completely randomized design, I used the BHH tomato example to teach the conduct and analysis of this design. (Anecdote. One of my proudest moments as a naive graduate student teaching stat methods came when an agriculture graduate student in my class told me he had successfully challenged his advisor’s experimental design for not having true replication.)

4.2 Example D. Metal Hardness Gauge Study

Montgomery’s paired experiment example pertains to hardness testing of metal specimens. I discussed this example in class, to facilitate the student’s understanding of the chapter but, as will be described, I selected one of Montgomery’s exercises as the primary vehicle for discussing the blocked two-treatment experiment. Montgomery’s example is more apt to be used by instructors and it raises some pertinent issues, so I will give my take on it here. In this example the testing machine measures hardness by pressing a rod with a pointed tip into the specimen and measuring the depth of the depression. The issue of interest is whether there is a difference in hardness readings obtained by using two different tips.

Two designs of the study are considered: One is to randomly select a number of specimens from those available, then randomly assign half of them to be hardness-tested with Tip 1 and half with Tip 2. It is known, though, that all specimens are not “exactly homogeneous,” thus hardness variability among specimens will tend to inflate the variability of hardness measurements by each tip, relative to the variability of multiple measurements on a single specimen. This design would thus make the comparison of tips more imprecise than one might like. Alternatively, it is decided, and it is feasible, to divide (split? mark?) each specimen into two parts and then test one part with Tip 1, the other with Tip 2, with random assignment of tips to parts and random ordering of testing. Interestingly, this example, as did BHH’s boys’ shoe example, pertains more to the choice of ex-

perimental unit—specimen or half-specimen?—than it does to the issue of whether a collection of existing experimental units can be usefully grouped into blocks.

In my introductory lectures, I say experimentation means active intervention in the cause system of a phenomenon; it means controlled manipulation of process inputs to see what happens to process output. The situation in this tip-testing example is a structured passive observation of a measurement process. Changing out the tips is not intervention. Thus, I would call this example a gauge study. Gauge studies, of course, are important, should be well-designed and -analyzed, and deserve to be taught. In most texts, numbers arrive on the page with no information about the measurement system. In practice, though, one needs evidence of a measurement system’s repeatability and reproducibility before undertaking experiments of interest. It must be admitted, though, that gauge studies have limited appeal to university students, even those in engineering and physical sciences, especially as their first exposure to designed experiments. We need to introduce experimental design with real attention-grabbers. (Anecdote: Bill Hunter told me that their editor wanted a title for their book (BHH) with sex appeal. Thus, *Statistics for Experimenters*, which is pretty subliminal but it’s there.) The BHH boys’ shoes example still works best for me.

4.3 Subject-Matter Passion

I tell my students that experiments are most successful if there is some subject-matter passion driving them. You don’t have to be a tomato grower or a shoe manufacturer, I think, to understand why some people could care deeply about such experiments. That’s another reason I like those examples. To show the importance of subject-matter passion I use an example that I call “Charlie Clark and the Car Charts.” The statistics group I managed at Sandia had a small library and when we bought a new book I routed it around to the group just so they would be aware of it. One such new book had to do with Graphical Methods. Charlie Clark was both thorough and a car nut. He did more than skim the table of contents. One chart he came across was a scatterplot of automobile engine displacement versus body weight. This plot showed a slightly curved positive association—heavier cars have bigger engines—and a couple of outlying points. The authors made the graphical point that you couldn’t “see” the relationship or the outliers in a table of the data and they commented that the outliers might be unusual cars or mistakes in the data. Then they went on to other topics.

The outlying points were two cars with unusually large engines for their body weights. They would be high-performance autos, so thorough Charlie not only looked at the chart, he got excited. He wanted one of those cars, so he looked up the source data (provided in the book’s appendixes). Alas, the two outliers were the Opel and Chevette, which he knew were performance dogs. He then went to the original *Consumer’s Reports* data source and found that the text authors had made transcription errors. The point I make in class is that Charlie found the true “message” in the data, which is what statistical analysis is all about, not because he was a better statistician than the authors, but because he had a passionate interest in the subject matter.

Many teachers of experimental design, including mine, Dave Weeks, circa 1965, require students to do projects and I have followed that practice (discussed in a later section). I encourage

students to do their project experiments in areas they are strongly interested in (as long as it is legal and safe). You can tell which students really care about how well paper airplanes fly and which ones are just going through the motions, so to speak.

4.4 Example E. Caliper Consistency

When I taught out of Montgomery (2001), instead of the metal-hardness gauge study, I used an exercise from the end of the chapter to introduce the paired experiment just because it seemed to be a more interesting set of data. It also turns out to raise some other pertinent issues, so I include it here.

This exercise is a study of two calipers used to measure ball bearings. The objective is to compare the calipers and the experimental design issue is (a) whether to have one set of technicians use one caliper and another set use the second or (b) whether each technician should use both calipers. The latter, of course, blocks out the technician-to-technician variability so we get a potentially more precise comparison of calipers.

This example is another gauge study. Further, it is a repeated measures study. There is only one experimental unit, a single ball bearing, in the study and it is measured twice by several inspectors, each using one caliper, then the other. The order is separately randomized for each inspector. I know that repeated measures designs can sometimes be properly analyzed via the same model used for randomized complete block designs, but I would rather put off that concept until later in the course.

These reservations occurred to me after I had used the caliper example in my University of Michigan class. The context I provided there was that the finding of no apparent difference between calipers was good; it meant that they could be used interchangeably (all manufacturers know that often some testers are kinder than others), at least for measuring one ball bearing size and material. There was a bonus finding here in that there also appeared to be no significant difference among technicians. Thus, it does not matter who is wielding the calipers. This operator-independence is another desirable property of a measurement system. Blocking “failed,” some would say, in this experiment because it did not remove a source of variation and thus improve the precision with which the calipers are compared, but from the point of demonstrating a good measurement system, the experiment was a success.

The usual paired t test or randomization analyses deal only with the differences between treatments, so differences among blocks are never considered. But there are often situations in which block differences are of interest and this example can be used to make that point. (In an ANOVA table, not yet introduced, the opportunity to evaluate block differences is always there if the question is pertinent and the test is valid. In this example, I used a data plot to conclude no differences among technicians.) I also make the point that in real-life one might want to expand this experiment to different ball bearings and multiple measurements by each technician. (My general point is that experiments are rarely one-time events, but rather steps in a sequential experimental program. Later in the course I illustrate this by experiments where, say, two treatments out of five appear to be winners, but there are not enough data to choose between them. The task then is to use what was learned in the present experiment to design a subsequent experiment that will

provide a more definitive comparison of the two contenders.) A real study of the measurement system would not be as small as this exercise. I must admit, though, that in an age of laser and other high-tech measurement of parts dimensions it is hard to generate much “passion” for studying a human-mechanical measurement system. Again, I’ll stick with the BHH shoe example for motivating the two-treatment blocked experiment.

5. EXPANDED EXAMPLE FROM SCOTT AND TRIGGS (2003)

Fundamentally, experiments consist of applying treatments to experimental units. Both treatments and experimental units can be “structured” according to one or more factors. In the case of multiple factors, the structure can be crossed or nested factors. This duality of treatments and experiments was driven home strongly to me by the structure of Triggs’ portion of the course notebook (Scott and Triggs 2003) that provided the organization of the course I taught at U Auckland. Triggs devotes separate chapters to “block structures,” which refers to the organization of experimental units, and to “treatment structures,” which refers to the single factor, or crossed or nested factors, used to define an experiment’s set of treatments. Only after this groundwork is laid does Triggs marry the two structures, along with the randomization scheme, to produce experimental designs. This approach contrasts markedly with the more conventional march through different experimental design families that most textbook authors take and that I had followed in previous courses.

In his chapter on block structures Triggs considers situations in which there is either (1) a single group of homogeneous eu’s, (2) multiple groups of eu’s, with the groups being defined by a single “blocking factor,” and (3) groups of eu’s defined by two blocking factors, either crossed or nested. Examples are given in which observations are made on these eu’s, but in none of the examples are the eu’s subjected to different treatments (in essence, they’re all “treated” the same). Triggs’ example of a single group of eu’s is strawberry yields by individual plants in a single garden with all the plants being treated identically (to a feasible extent). For a one-factor grouping of eu’s I made up an example of strawberry plant yields in multiple gardens. If we regard an experiment as an application of different treatments to eu’s, then these examples cannot be called experiments. They are (structured) observational studies. The analysis objective is to find out whether the structural factors influence the response and to characterize the associated fixed and random sources of variability. The story I construct for all of this is that at some future point we might consider an experiment with different treatments, such as fertilizers or insecticides. These strawberry studies are preliminary studies aimed at characterizing patterns of variability in the eu’s, between and within gardens, and thus would help us size and design future experiments.

5.1 Example F. Cloth Strength

The Scott and Triggs (2003) Lecture Notes are not generally available, so there is not much reason to describe the expanded contexts that an instructor might develop specifically for them. However, Triggs’ block-structure focus is unique, in my experience, and helpful in understanding the fundamentals of experimental design. Thus, I will use one of his examples to illustrate this approach and as a lead-in to a further discussion of blocks.

Cloth Strength, p. 91

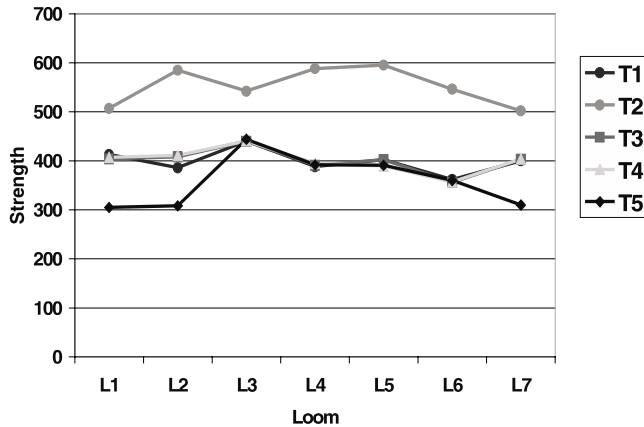


Figure 2. Average cloth strength by loom and technician.

Triggs' example of a two-factor block structure concerns textile production. Cloth can be produced on different looms by different technicians. A study is undertaken in which cloth samples are produced by each of five technicians using each of seven looms. The two factors are thus crossed and the five technicians and seven looms are randomly selected from large populations of each. Each technician produces one cloth sample on each loom so the result is 35 cloth samples (eu's) cross-classified by loom and technician. No treatments, at this point, are applied to these 35 eu's. The response of interest is cloth strength and Triggs' random-effects analysis estimates the loom and operator variance components. One might regard this as a uniformity study, similar in concept to the gauge studies discussed above.

My alternative (fixed-effects) analysis focused on these five technicians and seven looms. A plot of the data, Figure 2—Analysis One—shows that Technician 2 produces cloth distinctly stronger (nearly 50% more) than the other technicians. There is also some visual evidence that Technician 5 came up short on three of the looms. The ANOVA, for a two-way classification additive model, substantiates these impressions. The difference among technicians is highly significant; and the difference among looms is marginally significant due primarily to the one technician who under-performed on three of the looms. Of course, because there is no direct measurement of pure error variance in this study, the visual impression of loom-technician interaction cannot be substantiated. (The Tukey (1949) one degree-of-freedom test for interaction, though, might be applied.)

A favorite quote I use in class is Archie Bunker, to his son-in-law: "Don't give me no statistics (sic), Meathead. I want facts!" If this cloth example is used only to illustrate the calculations for a two-way ANOVA table or variance component estimation, we're just doing "statistics." The facts to communicate and to act on are the data patterns behind these numerical results, revealed in Figure 2. Thus, the further story line I provide is to interpret these results and discuss further actions. Technician 2 is either inadvertently making cloth stronger than required, and perhaps distorting some other property of the cloth, or else the other four technicians are all doing something wrong. Management needs to find out and correct the situation. Sound, data-based decisions and actions; that is our goal. It is the statistician's job to instigate such action. Before that, it is the statistician's

job to ferret out the message in the data. And before that it's the statistician's job to assure that the experiment will lead to data that are capable of revealing a message. If all a person did was carry out a two-way ANOVA and estimate the variance components without looking further at the data, the message (about the aberrance or excellence of Technician 2) would likely be missed.

Now, under the example's assumption that the five technicians were a random sample from a large workforce, we could conclude that (very) roughly 20% of this population is doing something unusual relative to the other 80%. (The variance component estimate does not convey this dichotomy.) Then, the course of action would be to instigate a workforce-wide study of what's going on and how it should be remedied. It's best to launch that effort, though, by finding out what makes the five technicians in this study operate the way they do. Also, I think a textile plant would want to know how specific looms are functioning and not just estimate their "population" variance. So my inclination, even when blocks are randomly selected, is to focus first on the information provided on the selected blocks and then consider broader implications, guided by subject-matter facts, not *statistics*.

6. BLOCKS

There are a variety of ways to think of blocks in experimental design. Triggs associates block structures with error structures, so block effects are always modeled as random effects. Each block term in the data model is assumed to be a random variable with a distribution characterized by its mean and variance. I find it difficult to take this general approach to blocks. For example, in medical experiments the experimental units, patients, can be blocked (grouped, in the usual sense in which the term is used) by sex, age, health, and other physical characteristics. The two sex levels, for example, are obviously not a random sample from some population of sexes. And sex is not a treatment that the experimenter applies to neutral experimental units. Sex is an inherent (structural) characteristic of the eu's to be used in the experiment, hence it is a (fixed) blocking factor. A reasonable analysis objective in this situation is to evaluate whether or not the two sexes respond to the treatments similarly and whether there is an overall difference between sexes in their responses to the treatments in the experiment. That is, it is reasonable to test for block by treatment interaction and for block differences in this situation, if possible, and to follow up and interpret the findings in terms of communicating and acting upon the observed treatment effects for both sexes.

Another view (e.g., Cobb 1998) is that blocking factors are "nuisance factors"—troublemakers that make it difficult to evaluate treatments unless these nuisances are properly corralled. Although that is true in some cases, it is decidedly not the case in the above examples. Experimental units grouped in meaningful blocks can improve the precision with which treatments are compared but blocks themselves can be of direct interest. Cobb (1998) apparently defined blocks in a more narrow sense than other authors, and I'm sure would not argue that sex is a nuisance factor, but I do not believe this narrow definition of blocking is helpful in light of the overwhelming use of the concept of blocking in a broader sense elsewhere.

Another important role of blocks is to determine the scope of the experiment: to set the boundaries within which subject-

matter considerations (not statistical sampling theory) would support inferences about how widely the observed treatment differences would obtain. BHH discuss how, in the shoe experiment, the experimenter ought to pick a wide variety of boys and conditions under which to compare the shoe sole materials. They don't say it's important to take a random sample of boys from a specific frame. Similarly, a fertilizer manufacturer would like to find out if one fertilizer outperforms its competitors in a variety of soils and growing environments, so running something like BHH's tomato fertilizer experiment at a (deliberately chosen) variety of locations spanning these conditions would be appropriate. Blocks selected to establish scope are not random and not nuisance factors. Defining blocks as inherent or created groups of eu's seems to me to be a unifying concept for the various roles that blocks play in designed experiments.

7. STUDENT PROJECTS

I have found that student projects provide my best vehicle for one-to-one communication with students, particularly in the large classes I taught at Michigan and Auckland, so I strongly advocate them, even at the expense of using class time that might be spent on lectures on additional design families or analysis methods. Projects also subject students to real-world pressure in determining a design, and in interpreting and communicating the results. This is the context that I had tried to evoke in my expanded textbook examples. Here are some details on how I have handled student projects.

At the University of New Mexico I asked each student to do a project on a topic of personal interest and importance. The process was that they draft an experimental plan that covered motivation, issues of interest, design, and anticipated analysis. This draft was then given to a fellow student for feedback and comment. I encouraged frankness and civility. The intent was to improve the quality of the plans and to have students experience the real-world roles of advocate and critic. Revised plans were then submitted to me for comment—not a grade at this point. I provided copious comments aimed generally at the aspects discussed above—experimental units, choice of treatments, blocking, replication, and randomization—and I suggested design changes, if warranted. All this is aimed at giving them experience in the collaborative, and sometimes combative, effort generally involved in determining an experimental design. They then ran their experiments, analyzed the data, and wrote reports. These I graded and again provided extensive written feedback. I required the graduate students in the class to give an oral report and gave the undergraduates the option of doing so also (with some bonus points attached).

At the University of Michigan, to encourage students to work collaboratively, my ground rules were that students could work in teams of one to four. Not inconsequentially, this also kept manageable the number of plans and final reports I would read and mark. We went through the same cycle of peer review, then instructor review of the written plans, followed by the experiment and written reports that I graded and commented on. Oral reports were again optional, for bonus points.

At the University of Auckland, because of the greater class size and reduced lecture hours, I only asked that student teams prepare experimental plans. (At Michigan I also had lab hours

that I used for some project activities.) Guidelines were for the experiment to have at least one blocking factor and at least two treatment factors. The experiments were to be feasible experiments that the students could do—for example, experiments to find a cure for cancer were beyond the scope. The plans went through peer feedback and my grading and extensive feedback.

In all three classes I encouraged students to use me as a consultant and this brought more students to my office than any other part of the class and provided great opportunities to talk about fundamental experimental design issues as they applied to their projects. I also devoted some lecture time to providing feedback on the plans—general areas of difficulty or strength. At Auckland I spiced this up by recognizing outstanding plans in several categories, including: most eu's; fewest eu's; and most subject-matter passion (two guys who you could tell really loved to fish and were determined to find the best location, depth, and bait).

8. CONCLUDING COMMENTS

As I wrote the first draft of this article I had just completed marking my final exams at U Auckland—an exercise that always provides an emotional roller coaster. One student can leave you euphoric—a kindred soul has been created. The next can leave you wondering why you even bothered. In between there are glimmers of varying luminosities (which I guess should not be unexpected; statistics is all about variation). Reading student evaluations takes you on another roller coaster. This has been true in all three university courses I have taught. Nevertheless, I remain committed to the proposition that we must teach more than formula plug-in in introductory experimental design and in all of our classes. Enough students resonate with this approach that I'm encouraged, not only about the approach but about the next generation. But I must admit that I am most motivated by my own conscience and convictions. (I'm sure that few instructors deliberately teach predominantly formula plug-in, but I'm also aware from personal experience that convenience and our own education and tendencies can lead an instructor down this path of least resistance.) I'm convinced that expanding textbook examples to include real-world-like contexts and provocative issues will better prepare students to intelligently use statistical experimental design for real. This approach shows respect for students and so should increase their respect for statistical thinking and methods.

I also am committed to teaching experimental design broadly. I think it is great to have a mixed group of students—statistics and other majors—in an experimental design class. In fact, I would like to form project teams with this mix for an even more realistic experience. These mixed classes have dual purposes which can make things difficult: (1) provide nonstatistics majors, who may never take another statistics class, with enough understanding of experimental design to function well in their subsequent coursework and careers; and (2) prepare statistics majors for more theoretical coursework and research. I would add a third purpose: (3) prepare statistics majors for collaborative work with professionals in other fields. My impression is that texts by statisticians and class syllabuses designed by statisticians emphasize (2). I would put more emphasis on (1) and (3). Statisticians suffer more by not understanding their collaborators than by not knowing enough theory.

One of my revered professors, the late Carl Marshall of Oklahoma State University, had a favorite statement that has stuck with me: “The nice thing about statistics is that the nouns may change but the verbs remain the same.” Let’s face it. We (statisticians) get our kicks out of the verbs—what we can do with data. It’s not surprising that our texts and teaching would emphasize the analysis aspects of experimental design. Our collaborators and clients, however, make their living on the nouns. If we teach only verbs (formula plug-in) to the next generation of statisticians and users of statistics, the connection between nouns and verbs will not be made. The facts will be obscured by Archie’s statistics.

How early we should teach experimental is another matter. Students who see how statistical experimental design can contribute to their anticipated careers are more receptive and engaged than are those who do not have that perspective and are just filling a slot on their transcript. I include both statistics and nonstatistics majors in that statement. It’s a matter of maturity and perspective, not mathematical ability. This career-view generally doesn’t come about until the senior or beginning graduate student level. Nevertheless, I believe Stat 101 should contain a healthy dose of experimental design. I would teach a unit based on the BHH tomato fertilizer and boys’ shoes experiments as a way to teach about experimental design, and life, not just paired and unpaired t test formulas.

I cannot claim (on sound statistical evidence) that my journey (literal and philosophical) aimed at improving the understanding and appreciation of the honorable profession of statistics among

the throngs who are exposed to it has been successful. I’ve benefited, though, by focusing on the essential features of statistical experimental design and struggling with how to engender understanding and enthusiasm for the topic among my students. I hope that my experiences and suggestions will benefit others engaged in the same struggle. Those who engage in this endeavor, with enthusiasm and creativity, for an entire career deserve our great respect. I greatly appreciate the three universities I have visited and look forward to further opportunities.

[Received October 2003. Revised April 2004.]

REFERENCES

- Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters*, New York: Wiley.
- Bross, I. D. (1974), “The Role of the Statistician: Scientist or Shoe Clerk,” *The American Statistician*, 28, 126–127.
- Cobb, G. W. (1998), *Introduction to Design and Analysis of Experiments*, New York: Springer-Verlag.
- Deming, W. E. (1975), “On Probability as a Basis for Action,” *The American Statistician*, 29, 146–152.
- Fisher, R. A. (1935a), *The Design of Experiments*, London: Oliver and Boyd.
- (1935b), *Statistical Methods for Research Workers*, London: Oliver and Boyd.
- Montgomery, D. C. (2001), *Design and Analysis of Experiments* (5th ed.), New York: John Wiley.
- Scott, A., and Triggs, C. (2003), Lecture Notes for Paper STATS 340, Dept. of Statistics, University of Auckland.
- Tukey, J. W. (1949), “One Degree Of Freedom For Non-additivity,” *Biometrics*, 5, 232.