4 The art of experimental design

Daniel Friedman and Alessandra Cassar

Your purpose determines the appropriate design for your experiment. It defines the *focus variables*, those whose effects you want to understand. But other things – the *nuisance variables* – may also have an effect and you need to account for them or you may reach incorrect conclusions.

For example, you might want to know what sorts of actions by others encourages a person to behave more altruistically, so your focus variable is others' actions. You should worry that altruistic behavior might also be affected by how you phrase instructions, so the wording of instructions is an important nuisance variable. On the other hand, if your purpose is to discover how phrasing can affect choices, then the wording of instructions is the focus and others' actions are an important nuisance. It all depends on your purpose.

The whole point of experimental design is to deal appropriately with both kinds of variables. You want the effects of your focus variables to come through sharply, and not be confounded with the effects of nuisances. There are two basic devices to separate out the effects, *control* and *randomization*. These complementary devices help to achieve *independence* (sometimes called design balance) among the variables affecting the outcomes.

This chapter is intended to help you understand these devices and the underlying ideas, so that you will know how to choose the most appropriate design for your purpose. The ideas turn out to be rather intuitive, but sometimes the terminology is a bit odd. Below you will see jargon like crossover and blocks, and in the wider literature you may encounter jargon like split-plot. These words hearken back to the roots of experimental design in agricultural experiments. The classic text is Fisher (1935), and we have found Box *et al.* (1978) very enlightening.

Control

You, the experimenter, can freely choose the values of many sorts of variables. For example, you can choose the institutions, say two different auction formats, and you can choose what sorts of cost to induce on the sellers. The deliberate choice, or control, of key variables is what distinguishes experimental data from happenstance data.

You basically have two options for controlling a variable. You can hold it *constant*, keeping it at the same level throughout the experiment. Or you can vary it between two or more levels, in which case it is called a *treatment* variable. For example, you can keep the same trading rules throughout your experiment, or you can have two different institutions like posted price and English auction. As you hold more variables constant, the experiment become simpler and cheaper, but you also learn less about the direct effects and the interactions among variables.

> Sometimes it takes some serious thinking and careful work through the theory before you can decide on the right control variable. Chapter 17 includes a prime example. The project is one of the few to consider nonlinear dynamics in the laboratory. Model predictions were given in terms of behavioral variables that should be observed (or inferred), so the control variables had to be picked indirectly.

Here are the standard rules of thumb on deciding which option to use:

1 Control all the variables that you can. It may be costly, but do not settle for happenstance unless you really have to.
2 Control your focus variables as treatments. Only by changing their level you can discover their effects.
3 For most treatments, two levels are sufficient for you to detect their effects. Detecting nonlinear effects requires more than two levels, but nonlinearity usually is not the main issue.
4 Separate the levels widely so that the effects will be evident.
5 Most nuisances should be controlled as constants, to economize on the design.
6 Nuisances that you think might interact with a focus variable, however, should be considered as treatments. An example of a possible interaction is a person might behave more altruistically after someone does him a favor when the instructions encourage altruism than when the instructions do not encourage altruism. In this case, an experimenter using the first sort of instructions would reach a different conclusion than one using the second sort. Both instructions should be used so that the interaction can be detected and incorporated in the conclusion.
7 Vary treatment variables *independently*.

Most of these rules are obvious once you think about them, but the last deserves further comment.

Independence

Treatment variables are *independent* if knowing the value of one variable does not give any information about the level of the other variables. The reason why you want to vary the treatment variables independently is simple. If two variables are dependent then their effects are harder (or impossible) to separate. Leamer (1983)

4 The art of experimental design

Daniel Friedman and Alessandra Cassar

Your purpose determines the appropriate design for your experiment. It defines the *focus variables*, those whose effects you want to understand. But other things – the *nuisance variables* – may also have an effect and you need to account for them or you may reach incorrect conclusions.

For example, you might want to know what sorts of actions by others encourages a person to behave more altruistically, so your focus variable is others' actions. You should worry that altruistic behavior might also be affected by how you phrase instructions, so the wording of instructions is an important nuisance variable. On the other hand, if your purpose is to discover how phrasing can affect choices, then the wording of instructions is the focus and others' actions are an important nuisance. It all depends on your purpose.

The whole point of experimental design is to deal appropriately with both kinds of variables. You want the effects of your focus variables to come through sharply, and not be confounded with the effects of nuisances. There are two basic devices to separate out the effects, *control* and *randomization*. These complementary devices help to achieve *independence* (sometimes called design balance) among the variables affecting the outcomes.

This chapter is intended to help you understand these devices and the underlying ideas, so that you will know how to choose the most appropriate design for your purpose. The ideas turn out to be rather intuitive, but sometimes the terminology is a bit odd. Below you will see jargon like crossover and blocks, and in the wider literature you may encounter jargon like split-plot. These words hearken back to the roots of experimental design in agricultural experiments. The classic text is Fisher (1935), and we have found Box *et al.* (1978) very enlightening.

Control

You, the experimenter, can freely choose the values of many sorts of variables. For example, you can choose the institutions, say two different auction formats, and you can choose what sorts of cost to induce on the sellers. The deliberate choice, or control, of key variables is what distinguishes experimental data from happenstance data.

You basically have two options for controlling a variable. You can hold it *constant*, keeping it at the same level throughout the experiment. Or you can vary it between two or more levels, in which case it is called a *treatment* variable. For example, you can keep the same trading rules throughout your experiment, or you can have two different institutions like posted price and English auction. As you hold more variables constant, the experiment become simpler and cheaper, but you also learn less about the direct effects and the interactions among variables.

> Sometimes it takes some serious thinking and careful work through the theory before you can decide on the right control variable. Chapter 17 includes a prime example. The project is one of the few to consider nonlinear dynamics in the laboratory. Model predictions were given in terms of behavioral variables that should be observed (or inferred), so the control variables had to be picked indirectly.

Here are the standard rules of thumb on deciding which option to use:

1 Control all the variables that you can. It may be costly, but do not settle for happenstance unless you really have to.
2 Control your focus variables as treatments. Only by changing their level you can discover their effects.
3 For most treatments, two levels are sufficient for you to detect their effects. Detecting nonlinear effects requires more than two levels, but nonlinearity usually is not the main issue.
4 Separate the levels widely so that the effects will be evident.
5 Most nuisances should be controlled as constants, to economize on the design.
6 Nuisances that you think might interact with a focus variable, however, should be considered as treatments. An example of a possible interaction is a person might behave more altruistically after someone does him a favor when the instructions encourage altruism than when the instructions do not encourage altruism. In this case, an experimenter using the first sort of instructions would reach a different conclusion than one using the second sort. Both instructions should be used so that the interaction can be detected and incorporated in the conclusion.
7 Vary treatment variables *independently*.

Most of these rules are obvious once you think about them, but the last deserves further comment.

Independence

Treatment variables are *independent* if knowing the value of one variable does not give any information about the level of the other variables. The reason why you want to vary the treatment variables independently is simple. If two variables are dependent then their effects are harder (or impossible) to separate. Leamer (1983) makes the point well in his satire of the Monetarist–Keynesian controversies of the 1970s. Leamer begins by supposing everyone accepts the fact that certain plants grow better under trees. One camp, the Aviophiles, argues that the cause is bird droppings, while another camp, the Luminists, argues for shade. Since shade and droppings are very highly correlated, the field data are inconclusive. A field experiment could settle the argument. Control the birds, say, by putting netting over some of the trees so that the two focus variables are independent. Then, you can check whether the plants grow as well under the dropping-free trees. Leamer's point is that experiments are more difficult to conduct for macroeconomic issues.

How do you make control variables independent? This is easy for variables you control as constants: they are trivially independent from all other variables. As for treatment variables, the first thing you might think of is to run *all conditions*, that is, all possible combinations of the treatment variables. For example, have equal acreage in each of the four conditions shade and droppings, no-shade and droppings, shade and no-droppings, and no-shade and no-droppings. We will see soon that there are more economical designs that also achieve independence.

Randomization

We are not yet out of the woods. The weather, or having an experiment late in the evening or during finals week, or something else might have an effect on your subjects' behavior. Some potential nuisance variables are not controllable, so independence seems problematic. For example, trees might grow more often on slopes facing north, and you do not have the time and money to change the landscape contours. The lack of control is especially serious when the nuisance variable is not even observable and may interact with a focus variable. For example, some subjects intrinsically are more altruistic than others in ways that are almost impossible to measure accurately. What if you happen to assign the more altruistic subjects to the first instruction conditions? Your conclusions might be completely wrong.

This insidious problem has an amazingly simple solution. About 80 years ago, the British statistician R.A. Fisher showed how randomization ensures independence. Assign the conditions in random order and your treatments will (eventually, as the number of trials increases) become independent of all uncontrolled variables, observable or not. For example, do not assign the first half of the subjects to arrive to the encouraging instructions; the first half may be intrinsically more (or less) altruistic. Instead, use a random device to choose which instructions to use for each subject. Randomization ensures independence as the number of subjects (or other random assignments) increases.

Efficient designs

Now that the principles are clear, let us go through some classical design schemes. For example, if your treatments are three institutions and two different subject pools, their combination gives you six conditions to cover. If you could run them simultaneously, there would not be any time issues. Usually, this is

impossible due to software limitations, different oral instructions, etc. So, no
you have to decide on the appropriate way to conduct the sessions. Here are som
of the options:

1 *Completely randomized*. In each session you draw the condition randoml
(with replacement) from the list of possible conditions. (That is, the chose
condition is an independent, identically distributed random variable with th
uniform distribution on the set of possible conditions.) This design is effec
tive, but it can become very expensive! In fact, by bad luck of the draw, you
budget might be exhausted before you run one of the conditions.

2 *Factorial*. This design is similar to the completely randomized design, excep
that the conditions are drawn without replacement until you exhaust you
finite number of copies (replications). For example, if you have six cond
tions to cover and you want to replicate them four times each, you need
"3 × 2 factorial design with four replications" that requires twenty-fou
sessions (run in *random* order, of course). This design allows you to neutraliz
the effects of nuisances that did not even occur to you as well as known b
uncontrollable nuisances.

Factorial design not only achieves complete independence among contr
variables in moderate numbers of trials, but also allows the examination
all the interactions. The design, however, has two disadvantages. First, th
number of conditions, hence the required number of trials, grows explosive
with increased number of treatment variables (or levels in each treatment
Second, it is not quite as robust to experimenter error as the fully randomize
design. If you make a mistake in assigning the treatments in one session, th
design is no longer factorial.

3 *Fractional factorial*. One way to decrease the required number of runs is
deliberately confound some treatments with high-order interactions that yo
believe are negligible. See Friedman and Sunder (1994) for a full explan
tion. This design allows you to reduce considerably the number of trials, b
it is less robust than a full factorial designs. Of course, if you make an err
in assigning treatments in this design, you can always revert to full factori
or randomized designs.

4 *Crossover*. You can run more than one condition in the same session. F
example, suppose you have a two-level treatment (A and B) and you ca
subdivide each session into four blocks (or sequences) of trials. Then you ca
run your first session with treatment sequence ABBA, the second sessio
with BAAB, and so on. This design can be economical. It is also conserv
tive in the sense that if the treatment effects linger, then the contrast observe
between the A and B runs will understate the true effect. See an example
a balanced design (variant on ABBA) in Chapter 18.

5 *Within-subjects* and *between-subjects*. Each subject sees all levels of
treatment variable in a within-subjects design. In the between-subjec
design, each subject just sees one level, but different subjects (possibly
different sessions) see different levels. As with the crossover design, th

makes the point well in his satire of the Monetarist–Keynesian controversies of the 1970s. Leamer begins by supposing everyone accepts the fact that certain plants grow better under trees. One camp, the Aviophiles, argues that the cause is bird droppings, while another camp, the Luminists, argues for shade. Since shade and droppings are very highly correlated, the field data are inconclusive. A field experiment could settle the argument. Control the birds, say, by putting netting over some of the trees so that the two focus variables are independent. Then, you can check whether the plants grow as well under the dropping-free trees. Leamer's point is that experiments are more difficult to conduct for macroeconomic issues.

How do you make control variables independent? This is easy for variables you control as constants: they are trivially independent from all other variables. As for treatment variables, the first thing you might think of is to run *all conditions*, that is, all possible combinations of the treatment variables. For example, have equal acreage in each of the four conditions shade and droppings, no-shade and droppings, shade and no-droppings, and no-shade and no-droppings. We will see soon that there are more economical designs that also achieve independence.

Randomization

We are not yet out of the woods. The weather, or having an experiment late in the evening or during finals week, or something else might have an effect on your subjects' behavior. Some potential nuisance variables are not controllable, so independence seems problematic. For example, trees might grow more often on slopes facing north, and you do not have the time and money to change the landscape contours. The lack of control is especially serious when the nuisance variable is not even observable and may interact with a focus variable. For example, some subjects intrinsically are more altruistic than others in ways that are almost impossible to measure accurately. What if you happen to assign the more altruistic subjects to the first instruction conditions? Your conclusions might be completely wrong.

This insidious problem has an amazingly simple solution. About 80 years ago, the British statistician R.A. Fisher showed how randomization ensures independence. Assign the conditions in random order and your treatments will (eventually, as the number of trials increases) become independent of all uncontrolled variables, observable or not. For example, do not assign the first half of the subjects to arrive to the encouraging instructions; the first half may be intrinsically more (or less) altruistic. Instead, use a random device to choose which instructions to use for each subject. Randomization ensures independence as the number of subjects (or other random assignments) increases.

Efficient designs

Now that the principles are clear, let us go through some classical design schemes. For example, if your treatments are three institutions and two different subject pools, their combination gives you six conditions to cover. If you could run them simultaneously, there would not be any time issues. Usually, this is impossible due to software limitations, different oral instructions, etc. So, now you have to decide on the appropriate way to conduct the sessions. Here are some of the options:

1 *Completely randomized*. In each session you draw the condition randomly (with replacement) from the list of possible conditions. (That is, the chosen condition is an independent, identically distributed random variable with the uniform distribution on the set of possible conditions.) This design is effective, but it can become very expensive! In fact, by bad luck of the draw, your budget might be exhausted before you run one of the conditions.

2 *Factorial*. This design is similar to the completely randomized design, except that the conditions are drawn without replacement until you exhaust your finite number of copies (replications). For example, if you have six conditions to cover and you want to replicate them four times each, you need a "3 × 2 factorial design with four replications" that requires twenty-four sessions (run in *random* order, of course). This design allows you to neutralize the effects of nuisances that did not even occur to you as well as known but uncontrollable nuisances.

Factorial design not only achieves complete independence among control variables in moderate numbers of trials, but also allows the examination of all the interactions. The design, however, has two disadvantages. First, the number of conditions, hence the required number of trials, grows explosively with increased number of treatment variables (or levels in each treatment). Second, it is not quite as robust to experimenter error as the fully randomized design. If you make a mistake in assigning the treatments in one session, the design is no longer factorial.

3 *Fractional factorial*. One way to decrease the required number of runs is to deliberately confound some treatments with high-order interactions that you believe are negligible. See Friedman and Sunder (1994) for a full explanation. This design allows you to reduce considerably the number of trials, but it is less robust than a full factorial designs. Of course, if you make an error in assigning treatments in this design, you can always revert to full factorial or randomized designs.

4 *Crossover*. You can run more than one condition in the same session. For example, suppose you have a two-level treatment (A and B) and you can subdivide each session into four blocks (or sequences) of trials. Then you can run your first session with treatment sequence ABBA, the second session with BAAB, and so on. This design can be economical. It is also conservative in the sense that if the treatment effects linger, then the contrast observed between the A and B runs will understate the true effect. See an example of a balanced design (variant on ABBA) in Chapter 18.

5 *Within-subjects* and *between-subjects*. Each subject sees all levels of a treatment variable in a within-subjects design. In the between-subjects design, each subject just sees one level, but different subjects (possibly in different sessions) see different levels. As with the crossover design, the

within-subjects design is conservative. But it controls for subjects' personal idiosyncrasies, which sometimes are an important nuisance.

6 *Matched pairs*. The idea of controlling for nuisances by varying only one treatment appears in its purest form in the matched pairs design. A classical example is the boys' shoes experiment testing the durability of a new shoe sole material. Instead of giving either the old or the new soles to different boys, each boy received one old and one new sole randomly assigned to the left or the right foot. In this way, nuisances such as the subject habits and level of activity are controlled.

As another example, consider the experiments that allowed Team New Zealand to win the 1995 America's Cup, ending the longest winning streak in sport history. (Team US had held the cup for 132 years!) Instead of building two different boats and testing the keels separately for each model, New Zealand built two virtually identical boats and tested different keel configurations by racing the two boats against each other, thus also controlling for weather and sea conditions. This design helped them improve at a rate of 20–30 s per month versus the traditional 7–15 s per month.

One clever way to get matched pairs in the laboratory is to have subjects make two decisions each trial for two environments that differ only in one treatment. For example, Kagel and Levin (1986) have subjects bid the same value draw in both a small group auction and a large group auction. More recently, Falk *et al.* (2003) use a similar dual trial design to isolate neighborhood effects.

Other, less classical designs sometimes are useful:

7 *Baseline neighborhood*. In this design, one picks a baseline condition (combination of treatment levels) and changes one treatment at a time. For example, one of the authors currently is investigating how ten different variables affect the strength of the sunk cost fallacy. Pilot experiments disclosed a baseline combination that seems strong. A factorial design is infeasible because even one replication with only two levels for each treatment would require $2^{10} = 1{,}024$ sessions. The plan is to just vary one treatment at a time (e.g. the instructions or the cost differential) from the baseline in crossover type run sequences for each subject. This design will not tell us much about how the variables interact, but it will give a first look at the main effects.

Important nuisances

In choosing your design it is worth thinking through what nuisance variables are likely to be important and how you will deal with them. Here is a checklist of standard nuisances.

1 *Learning*. Subjects' behavior usually changes over time as their understanding of game deepens during a session. If this is a nuisance, you can control it by keeping it constant: use only the last few periods or runs. You can control it as a treatment too, by using a balanced crossover design.

2 *Experience*. This problem is similar to learning, but occurs across sessions. To avoid it, it is good practice to keep the experience of the subjects under control. Keep a database to track which subjects already came and played in a particular experiment. The easiest solution is to use only inexperienced subjects, but often you want to confirm the results with experienced subjects. Unless it is part of your research question, do not mix experienced and inexperienced subjects in the same market or game session.

3 *Boredom and fatigue*. Try to keep your sessions no more than 2 h (unless required by your treatment), and shorter is even better. You may save some money and time by running fewer but longer sessions, but you may pay too high a price. Salience and dominance are compromised when your data come from tired or bored subjects.

4 *Extracurricular contact*. Pay attention and try to prevent any uncontrolled communication among your subjects during a session. During a restroom break, they may decide to collude! So, if you cannot monitor them, change the parameters after each break; this will thwart most collusion attempts.

5 *Self-selection*. Try to have a long list from which you can choose your subjects. When the subject pool is potentially important, you should actively choose balanced subject pools. For example, if you advertise a finance experiment in a finance class and in a biology class and let them show up at the door, you probably will end up mainly with finance students.

6 *Idiosyncrasies of individual subjects or pools*. A subject or a group with a particular background may lead to unrepresentative behavior. We had once a scheduler that was member of a sorority, and after a couple of sessions we realized our subjects were exclusively first year female members! Try to avoid these obvious occurrences, and replicate with different pools to take care of phenomena not so visible.

We conclude with a final piece of advice.

KISS: keep it simple! More elaborate experimental designs usually cause more problems for beginners than they solve.

References

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978) *Statistics for Experimenters*, New York: John Wiley and Sons.

Falk, A., Fischbacher, U., and Gächter, S. (2003) "Living in two neighborhoods – social interactions in the lab," No iewwp150 in IEW – Working Papers from Institute for Empirical Research in Economics – IEW.

Fisher, R.A. (1935) *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Friedman, D. and Sunder S. (1994) *Experimental Methods: A Primer for Economists*, Cambridge: Cambridge University Press.

Kagel, J.H. and Levin, D. (1986) "The winner's curse and public information in common value auctions," *American Economic Review*, 76: 894–920.

Leamer, E. (1983) "Lets take the con out of econometrics," *American Economic Review*, 73: 31–43.

within-subjects design is conservative. But it controls for subjects' personal idiosyncrasies, which sometimes are an important nuisance.

6 *Matched pairs*. The idea of controlling for nuisances by varying only one treatment appears in its purest form in the matched pairs design. A classical example is the boys' shoes experiment testing the durability of a new shoe sole material. Instead of giving either the old or the new soles to different boys, each boy received one old and one new sole randomly assigned to the left or the right foot. In this way, nuisances such as the subject habits and level of activity are controlled.

As another example, consider the experiments that allowed Team New Zealand to win the 1995 America's Cup, ending the longest winning streak in sport history. (Team US had held the cup for 132 years!) Instead of building two different boats and testing the keels separately for each model, New Zealand built two virtually identical boats and tested different keel configurations by racing the two boats against each other, thus also controlling for weather and sea conditions. This design helped them improve at a rate of 20–30 s per month versus the traditional 7–15 s per month.

One clever way to get matched pairs in the laboratory is to have subjects make two decisions each trial for two environments that differ only in one treatment. For example, Kagel and Levin (1986) have subjects bid the same value draw in both a small group auction and a large group auction. More recently, Falk *et al.* (2003) use a similar dual trial design to isolate neighborhood effects.

Other, less classical designs sometimes are useful:

7 *Baseline neighborhood*. In this design, one picks a baseline condition (combination of treatment levels) and changes one treatment at a time. For example, one of the authors currently is investigating how ten different variables affect the strength of the sunk cost fallacy. Pilot experiments disclosed a baseline combination that seems strong. A factorial design is infeasible because even one replication with only two levels for each treatment would require $2^{10} = 1{,}024$ sessions. The plan is to just vary one treatment at a time (e.g. the instructions or the cost differential) from the baseline in crossover type run sequences for each subject. This design will not tell us much about how the variables interact, but it will give a first look at the main effects.

Important nuisances

In choosing your design it is worth thinking through what nuisance variables are likely to be important and how you will deal with them. Here is a checklist of standard nuisances.

1 *Learning*. Subjects' behavior usually changes over time as their understanding of game deepens during a session. If this is a nuisance, you can control it by keeping it constant: use only the last few periods or runs. You can control it as a treatment too, by using a balanced crossover design.

2 *Experience*. This problem is similar to learning, but occurs across sessions. To avoid it, it is good practice to keep the experience of the subjects under control. Keep a database to track which subjects already came and played in a particular experiment. The easiest solution is to use only inexperienced subjects, but often you want to confirm the results with experienced subjects. Unless it is part of your research question, do not mix experienced and inexperienced subjects in the same market or game session.

3 *Boredom and fatigue*. Try to keep your sessions no more than 2 h (unless required by your treatment), and shorter is even better. You may save some money and time by running fewer but longer sessions, but you may pay too high a price. Salience and dominance are compromised when your data come from tired or bored subjects.

4 *Extracurricular contact*. Pay attention and try to prevent any uncontrolled communication among your subjects during a session. During a restroom break, they may decide to collude! So, if you cannot monitor them, change the parameters after each break; this will thwart most collusion attempts.

5 *Self-selection*. Try to have a long list from which you can choose your subjects. When the subject pool is potentially important, you should actively choose balanced subject pools. For example, if you advertise a finance experiment in a finance class and in a biology class and let them show up at the door, you probably will end up mainly with finance students.

6 *Idiosyncrasies of individual subjects or pools*. A subject or a group with a particular background may lead to unrepresentative behavior. We had once a scheduler that was member of a sorority, and after a couple of sessions we realized our subjects were exclusively first year female members! Try to avoid these obvious occurrences, and replicate with different pools to take care of phenomena not so visible.

We conclude with a final piece of advice.

KISS: keep it simple! More elaborate experimental designs usually cause more problems for beginners than they solve.

References

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978) *Statistics for Experimenters*, New York: John Wiley and Sons.

Falk, A., Fischbacher, U., and Gächter, S. (2003) "Living in two neighborhoods – social interactions in the lab," No iewwp150 in IEW – Working Papers from Institute for Empirical Research in Economics – IEW.

Fisher, R.A. (1935) *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Friedman, D. and Sunder S. (1994) *Experimental Methods: A Primer for Economists*, Cambridge: Cambridge University Press.

Kagel, J.H. and Levin, D. (1986) "The winner's curse and public information in common value auctions," *American Economic Review*, 76: 894–920.

Leamer, E. (1983) "Lets take the con out of econometrics," *American Economic Review*, 73: 31–43.

> omists. . . . But even if our Hayek hypothesis continues to outperform its competitors in laboratory experiments, does this mean it will do comparably well in the "field" environment of the economy? On the assumption of parallelism, namely that the same physical (and behavioral) laws hold everywhere, it is a reasonable working hypothesis, provisionally, to make this extension, but independent field observations, or experiments, are the appropriate vehicle for testing the extended hypothesis. (1982a, p. 177).

Gode and Sunder (1992, 1993a,b) illustrate the fruitful interplay between experiment and computer simulation, and add a new twist on the Hayek hypothesis. The authors create zero-intelligence (ZI) computerized traders that bid or ask randomly subject to a no-loss constraint. They find that the double-auction institution produces highly efficient outcomes even with ZI traders! Perhaps the rationality assumption plays a smaller role in some market institutions than most economists have presumed.

3

Experimental design

How does the number of buyers and sellers affect market efficiency? Do consumers prefer the "new improved" product or the "classic" version? Whether your purposes are scientific or commercial, you probably are interested in the effects of only a few variables, the *focus* variables. Usually you must also keep track of several other variables of little or no direct interest, the *nuisance* variables, because they may affect your results.

Which variables are focus and which are nuisance in your experiment depends on your purpose. The number of buyers is a focus variable in some oligopoly experiments, but the same variable is a nuisance in experiments testing consumer response to new products.

This chapter will explain how to design experiments that sharpen the effects of focus variables and minimize blurring due to nuisance variables. It will also explain how to design experiments that allow you to disentangle the effects of different variables, that is, how to avoid *confounding* the effects of two or more variables.

The first two sections introduce control and randomization, the basic ingredients of proper experimental design. Sections 3.3 and 3.4 elaborate on these ingredients and discuss specific designs. Distilled practical advice appears in the next section, and the last section illustrates the main ideas while reviewing some "test-bed" market experiments.

A word of warning before we begin. This chapter contains technical jargon. We have tried to follow the most common practices, but the literature is not entirely consistent in how words are used. You can consult the glossary at the end of the book to see how we use these words, but be careful in reading the literature to check what the author really means.

3.1 Direct experimental control: Constants and treatments

In the laboratory you can directly control many variables. You can freely select cost and value parameters and trading rules in market experiments, or the choice set and the subject pool in individual choice experiments. By controlling important variables you produce experimental data rather than happenstance data.

The simplest way to control a variable is to hold it *constant* at some convenient level. For example, enforce the same double-auction trading rules throughout a market experiment. The main alternative is to chose two or more different levels that may produce sharply different outcomes, and to control the variable at each chosen level for part of the experiment (or subset of experiments). For example, use two different sets of cost parameters, one inducing highly elastic supply and the other inelastic supply. Perhaps because of their prevalence in medical experiments, variables controlled at two or more levels are called *treatment* variables.

There is a tradeoff between controlling variables as constants and as treatments. As you hold more variables constant your experiment becomes simpler and cheaper, but you learn less about the direct effects and the interactions among the variables. Section 3.5 offers some suggestions on managing this tradeoff.

Suppose you choose two treatment variables, say the market institution with levels PO (posted offer) and DA (double auction), and the demand elasticity with levels E (elastic) and I (inelastic). Despite your control, you will completely confound their effects if you always change the variables together, say PO-E combination half the time and DA-I combination the other half. Instead, if you run each treatment combination (PO-E, PO-I, DA-E, and DA-I) one quarter of the time, you can gauge the separate effects of the two treatments. The logic is quite general: *Vary all treatment variables independently* to obtain the clearest possible evidence on their effects (see Figure 3.1).

3.2 Indirect control: Randomization

Some variables are difficult or impossible to control. For example, weather is an important and uncontrollable nuisance in agricultural experiments. (And occasionally in economic experiments: One of the authors recalls snowstorms preventing subjects from showing up and the other author remembers watching helplessly as airconditioning failed and the room temperature rose above 100°F in an early computerized experiment.) For economists, subjects' expectations usually are more important than the weather and just as uncontrollable. Some potentially

A. Confounded Treatment Variables:

	Elastic Demand	Inelastic Demand
Posted Offer Auction	Observations (PO-E)	No Observations
Double Auction	No Observations	Observations (DA-I)

B. Independent Treatment Variables:

	Elastic Demand	Inelastic Demand
Posted Offer Auction	Observations (PO-E)	Observations (PO-I)
Double Auction	Observations (DA-E)	Observations (DA-I)

Fig. 3.1 Independent variation of treatment variables.

important nuisances, such as a subject's alertness and interest, are not even *observable* by the experimenter, much less controllable.

Uncontrolled nuisances can cause inferential errors if they are confounded with focus variables. The real cause of improvement in harvests in the year a new seed variety is introduced may be good weather. Efficiency may decline when elastic supply parameters are introduced late in a long experiment, but the reason may be subjects' fatigue. The problem is that you may attribute an observed effect to a focus variable when the effect actually arises from an uncontrolled nuisance.

How can you avoid confounding problems when you can't directly control some important nuisances? The advice offered at the end of the previous section provides a hint. Independence among controlled var-

3.1 **Direct experimental control: Constants and treatments**

In the laboratory you can directly control many variables. You can freely select cost and value parameters and trading rules in market experiments, or the choice set and the subject pool in individual choice experiments. By controlling important variables you produce experimental data rather than happenstance data.

The simplest way to control a variable is to hold it *constant* at some convenient level. For example, enforce the same double-auction trading rules throughout a market experiment. The main alternative is to chose two or more different levels that may produce sharply different outcomes, and to control the variable at each chosen level for part of the experiment (or subset of experiments). For example, use two different sets of cost parameters, one inducing highly elastic supply and the other inelastic supply. Perhaps because of their prevalence in medical experiments, variables controlled at two or more levels are called *treatment* variables.

There is a tradeoff between controlling variables as constants and as treatments. As you hold more variables constant your experiment becomes simpler and cheaper, but you learn less about the direct effects and the interactions among the variables. Section 3.5 offers some suggestions on managing this tradeoff.

Suppose you choose two treatment variables, say the market institution with levels PO (posted offer) and DA (double auction), and the demand elasticity with levels E (elastic) and I (inelastic). Despite your control, you will completely confound their effects if you always change the variables together, say PO-E combination half the time and DA-I combination the other half. Instead, if you run each treatment combination (PO-E, PO-I, DA-E, and DA-I) one quarter of the time, you can gauge the separate effects of the two treatments. The logic is quite general: *Vary all treatment variables independently* to obtain the clearest possible evidence on their effects (see Figure 3.1).

3.2 **Indirect control: Randomization**

Some variables are difficult or impossible to control. For example, weather is an important and uncontrollable nuisance in agricultural experiments. (And occasionally in economic experiments: One of the authors recalls snowstorms preventing subjects from showing up and the other author remembers watching helplessly as airconditioning failed and the room temperature rose above 100°F in an early computerized experiment.) For economists, subjects' expectations usually are more important than the weather and just as uncontrollable. Some potentially

A. Confounded Treatment Variables:

	Elastic Demand	Inelastic Demand
Posted Offer Auction	Observations (PO-E)	No Observations
Double Auction	No Observations	Observations (DA-I)

B. Independent Treatment Variables:

	Elastic Demand	Inelastic Demand
Posted Offer Auction	Observations (PO-E)	Observations (PO-I)
Double Auction	Observations (DA-E)	Observations (DA-I)

Fig. 3.1 Independent variation of treatment variables.

important nuisances, such as a subject's alertness and interest, are not even *observable* by the experimenter, much less controllable.

Uncontrolled nuisances can cause inferential errors if they are confounded with focus variables. The real cause of improvement in harvests in the year a new seed variety is introduced may be good weather. Efficiency may decline when elastic supply parameters are introduced late in a long experiment, but the reason may be subjects' fatigue. The problem is that you may attribute an observed effect to a focus variable when the effect actually arises from an uncontrolled nuisance.

How can you avoid confounding problems when you can't directly control some important nuisances? The advice offered at the end of the previous section provides a hint. Independence among controlled variables prevents confounding problems. We would solve the present problem if we could somehow make the uncontrolled nuisances independent of the treatment variables.

Randomization provides indirect control of uncontrolled (even unobservable) variables by ensuring their *eventual* independence of treatment variables. The basic idea is to assign chosen levels of the treatment variables in random order. For example, in a market experiment subjects' personal idiosyncracies and habits are an uncontrollable and largely unobservable nuisance variable. When subjects arrive, don't assign all the early birds to the role of sellers and the late arrivals to the role of buyers. Randomize the assignment and you can be confident that observed profit differences between buyers and sellers arise from differences in the roles and not from differences in subjects' personal characteristics.

The simplest valid experimental design is called *completely randomized.* In this design, each treatment (or each conjunction of treatment variables) is equally likely to be assigned in each trial. (A *trial* is an indivisible unit of an experiment, such as a trading period in a market experiment.) Suppose you choose a completely randomized design for the two-treatment experiment illustrated in Figure 3.1. Then in each trial you might flip two fair coins to select each of the four treatments PO-E, PO-I, DA-E, and DA-I with probability 0.25 in each trial, independently of selections in previous trials.

Complete randomization is quite effective when you can afford to run many trials. Independence among your treatment variables and uncontrolled nuisance variables is "eventual" in the sense that only as the number of trials gets arbitrarily large does the probability of a given positive or negative correlation go to zero. You can occasionally get a large correlation between treatments and uncontrolled nuisances in a small set of randomized trials. Classical statistical techniques, discussed in Chapter 7, take this problem into account.

When uncontrolled nuisances produce little variation across trials, the completely randomized design is hard to improve upon. When controllable nuisances do significantly affect outcomes, however, designs that appropriately combine control with randomization are more efficient in the sense that they can produce equally decisive results from fewer trials. These designs ensure zero correlation among controlled variables even in small sets of trials.

Random block is the general name given to this improved design. The difference from the completely randomized design is that one or more nuisance variables are controlled as treatments rather than randomized.

Nuisance treatment variables are often called blocking variables, held constant within a block [subset of trials] but varied across blocks. The next two subsections provide examples.

3.3 The within-subjects design as an example of blocking and randomization

The purpose of the classic boys' shoe experiment (Box, Hunter, and Hunter, 1978, p. 97ff) is to see whether a new sole material lasts longer than the old. The focus is sole material, a treatment variable with two levels: old and new. Measured wear varies considerably, mostly from subjects' different activities and habits: Some boys are couch potatoes, others ride scooters using a shoe for a brake. Clever experimental design prevents these nuisances from obscuring the focus variable's effects: Each boy gets a pair of shoes with one sole of new material and the other sole of old. Thus subject identity in this design is a blocking (i.e., nuisance treatment) variable that captures the habits and activities nuisances, and *differences* in measured wear between left and right soles becomes the relevant performance measure. Random assignment of the focus variable (new material on left or right shoe) reduces confounding due to other nuisances, such as whether scooter brakers tend to be left or right footed.

Experimental designs that vary levels of the focus variable only across subjects are generically called *between subjects* designs and those that use several different levels for each subject are called *within-subjects* designs. The shoe experiment uses a special within-subjects design that allows all data to be expressed as differences across matched pairs. The matched-pair differences allow sharper inferences to the extent that individual subject variation is an important nuisance.

The same trick can be useful in economics experiments. For example, suppose you conduct individual choice experiments comparing the willingness to pay (WTP) for a gamble to the willingness to accept (WTA) a certain payment in lieu of the gamble. If you want to see whether your new "transparent" instructions will bring WTP and WTA closer together, then individual variability is an important nuisance you should take into account – for instance, some subjects may be more risk averse than others and report low WTP and low WTA. It would be appropriate to employ a within-subjects design as in the shoe experiment. Specifically, you could ask each subject for WTPs and WTAs in random order, and analyze the *differences* WTA – WTP across subjects for each gamble. That way you eliminate a potentially important source of noise, and the effects of your focus (instructions) then become more visible.

iables prevents confounding problems. We would solve the present problem if we could somehow make the uncontrolled nuisances independent of the treatment variables.

Randomization provides indirect control of uncontrolled (even unobservable) variables by ensuring their *eventual* independence of treatment variables. The basic idea is to assign chosen levels of the treatment variables in random order. For example, in a market experiment subjects' personal idiosyncracies and habits are an uncontrollable and largely unobservable nuisance variable. When subjects arrive, don't assign all the early birds to the role of sellers and the late arrivals to the role of buyers. Randomize the assignment and you can be confident that observed profit differences between buyers and sellers arise from differences in the roles and not from differences in subjects' personal characteristics.

The simplest valid experimental design is called *completely randomized.* In this design, each treatment (or each conjunction of treatment variables) is equally likely to be assigned in each trial. (A *trial* is an indivisible unit of an experiment, such as a trading period in a market experiment.) Suppose you choose a completely randomized design for the two-treatment experiment illustrated in Figure 3.1. Then in each trial you might flip two fair coins to select each of the four treatments PO-E, PO-I, DA-E, and DA-I with probability 0.25 in each trial, independently of selections in previous trials.

Complete randomization is quite effective when you can afford to run many trials. Independence among your treatment variables and uncontrolled nuisance variables is "eventual" in the sense that only as the number of trials gets arbitrarily large does the probability of a given positive or negative correlation go to zero. You can occasionally get a large correlation between treatments and uncontrolled nuisances in a small set of randomized trials. Classical statistical techniques, discussed in Chapter 7, take this problem into account.

When uncontrolled nuisances produce little variation across trials, the completely randomized design is hard to improve upon. When controllable nuisances do significantly affect outcomes, however, designs that appropriately combine control with randomization are more efficient in the sense that they can produce equally decisive results from fewer trials. These designs ensure zero correlation among controlled variables even in small sets of trials.

Random block is the general name given to this improved design. The difference from the completely randomized design is that one or more nuisance variables are controlled as treatments rather than randomized.

Nuisance treatment variables are often called blocking variables, held constant within a block [subset of trials] but varied across blocks. The next two subsections provide examples.

3.3 The within-subjects design as an example of blocking and randomization

The purpose of the classic boys' shoe experiment (Box, Hunter, and Hunter, 1978, p. 97ff) is to see whether a new sole material lasts longer than the old. The focus is sole material, a treatment variable with two levels: old and new. Measured wear varies considerably, mostly from subjects' different activities and habits: Some boys are couch potatoes, others ride scooters using a shoe for a brake. Clever experimental design prevents these nuisances from obscuring the focus variable's effects: Each boy gets a pair of shoes with one sole of new material and the other sole of old. Thus subject identity in this design is a blocking (i.e., nuisance treatment) variable that captures the habits and activities nuisances, and *differences* in measured wear between left and right soles becomes the relevant performance measure. Random assignment of the focus variable (new material on left or right shoe) reduces confounding due to other nuisances, such as whether scooter brakers tend to be left or right footed.

Experimental designs that vary levels of the focus variable only across subjects are generically called *between subjects* designs and those that use several different levels for each subject are called *within-subjects* designs. The shoe experiment uses a special within-subjects design that allows all data to be expressed as differences across matched pairs. The matched-pair differences allow sharper inferences to the extent that individual subject variation is an important nuisance.

The same trick can be useful in economics experiments. For example, suppose you conduct individual choice experiments comparing the willingness to pay (WTP) for a gamble to the willingness to accept (WTA) a certain payment in lieu of the gamble. If you want to see whether your new "transparent" instructions will bring WTP and WTA closer together, then individual variability is an important nuisance you should take into account – for instance, some subjects may be more risk averse than others and report low WTP and low WTA. It would be appropriate to employ a within-subjects design as in the shoe experiment. Specifically, you could ask each subject for WTPs and WTAs in random order, and analyze the *differences* WTA − WTP across subjects for each gamble. That way you eliminate a potentially important source of noise, and the effects of your focus (instructions) then become more visible.

3.4 Other efficient designs

The within-subjects idea has two useful variants. A *crossover* design takes a subject or group of subjects and varies the levels, say A and B, of a treatment variable across trials. When you suspect your treatment variable has effects lasting several trials, you should consider the ABA crossover design. (The simpler AB design confounds time and learning with the treatment variable.) For example, suppose your focus variable is the market institution with A = the double auction and B = buyers' auction (sellers passive). The convergence behavior of a group of traders may carry over from one trading period to the next, so in one session you might conduct four A trading periods followed by eight B trading periods and finish with four more A periods (ABA), and use the complementary BAB design in a companion session. Then the difference in mean observed performance between the A and B periods would conservatively indicate the effect of your focus variable.

A second variant, the *dual trial,* is especially useful when individual or group idiosyncrasies may be an important nuisance. Kagel and Levin (1986), for example, suspected that individual random signals and the behavior of other bidders in a group could affect bidder behavior in first-price common-values auctions. To test cleanly the effects of the focus variable, group size with levels *S*(mall) and *L*(arge), they employed dual auctions: upon receiving her signal, each subject submitted two bids, one for a small-group auction and a second for the large-group auction. Their dual auction design allowed the authors to isolate the effect of group size by looking at differences ($b_L - b_S$) in the two bids across subjects and time periods.

The *factorial design* is perhaps the most important general method for combining randomization and direct control when you have two or more treatment variables. To illustrate, consider two treatment variables ("factors") labeled *R* and *S,* with three levels H(igh), M(edium) and L(ow) for *R* and two levels H(igh) and L(ow) for *S.* In the resulting 3 × 2 factorial design, each of the six treatments LL, LH, ML, MH, HL, and HH is employed in the same number *k* of trials. Thus $3 \times 2 \times 4 = 24$ trials are required to replicate the design $k = 4$ times. Randomization plays an essential role in that you must assign the six treatments in random order to the six trials in each replication.

When it is feasible, the factorial design is more efficient than the completely randomized design because it ensures that each treatment (combination) occurs an equal number *k* of times, and that the treatment variables all have zero correlations even for small replication numbers *k.* Among other things, this helps you to distinguish the direct effects of the treatment variables from interactions.

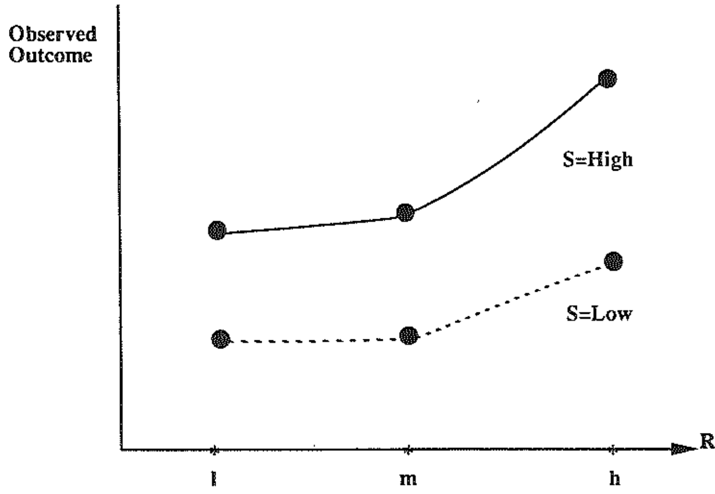

Fig. 3.2 Mean outcomes in a hypothetical factorial experiment.

Figure 3.2 uses the 3×2 example to illustrate direct and interacti effects. The vertical axis is the observed outcome, say market efficienc The first treatment variable *R,* say elasticity of demand and suppl appears on the horizontal axis and the second variable *S,* say payo intensity, shows up in the two curves labeled S = High and S = Lo The curves themselves connect the hypothetical mean outcomes in ea treatment. The distance between the curves measures the direct effe of variable *S.* When the curves are parallel, there is no interactio between *R* and *S,* but when the gap between the curves widens as Figure 3.1, there is a positive *RS* interaction. Chapter 7 will discuss t issue more extensively.

The factorial design is a bit less robust than the fully randomiz design because experimenter errors in assigning treatments and missi trials (from computer glitches or no-show subjects, for instance) mo seriously impair the data analysis. Indeed, if these problems are fr quent, the factorial design becomes indistinguishable from the con pletely randomized.

Another problem with the basic factorial design is that the numb of required trials increases quickly as the number of factors increase Suppose, for example, you chose only two levels for each treatme variable. Even then, you need $2^4 = 16$ trials for 4 factors and $2^8 = 2$ trials for eight factors to run just a single replication! The problem

3.4 **Other efficient designs**

The within-subjects idea has two useful variants. A *crossover* design takes a subject or group of subjects and varies the levels, say A and B, of a treatment variable across trials. When you suspect your treatment variable has effects lasting several trials, you should consider the ABA crossover design. (The simpler AB design confounds time and learning with the treatment variable.) For example, suppose your focus variable is the market institution with A = the double auction and B = buyers' auction (sellers passive). The convergence behavior of a group of traders may carry over from one trading period to the next, so in one session you might conduct four A trading periods followed by eight B trading periods and finish with four more A periods (ABA), and use the complementary BAB design in a companion session. Then the difference in mean observed performance between the A and B periods would conservatively indicate the effect of your focus variable.

A second variant, the *dual trial,* is especially useful when individual or group idiosyncrasies may be an important nuisance. Kagel and Levin (1986), for example, suspected that individual random signals and the behavior of other bidders in a group could affect bidder behavior in first-price common-values auctions. To test cleanly the effects of the focus variable, group size with levels *S*(mall) and *L*(arge), they employed dual auctions: upon receiving her signal, each subject submitted two bids, one for a small-group auction and a second for the large-group auction. Their dual auction design allowed the authors to isolate the effect of group size by looking at differences $(b_L - b_S)$ in the two bids across subjects and time periods.

The *factorial design* is perhaps the most important general method for combining randomization and direct control when you have two or more treatment variables. To illustrate, consider two treatment variables ("factors") labeled *R* and *S,* with three levels H(igh), M(edium) and L(ow) for *R* and two levels H(igh) and L(ow) for *S.* In the resulting 3 × 2 factorial design, each of the six treatments LL, LH, ML, MH, HL, and HH is employed in the same number *k* of trials. Thus $3 \times 2 \times 4 = 24$ trials are required to replicate the design $k = 4$ times. Randomization plays an essential role in that you must assign the six treatments in random order to the six trials in each replication.

When it is feasible, the factorial design is more efficient than the completely randomized design because it ensures that each treatment (combination) occurs an equal number *k* of times, and that the treatment variables all have zero correlations even for small replication numbers *k*. Among other things, this helps you to distinguish the direct effects of the treatment variables from interactions.

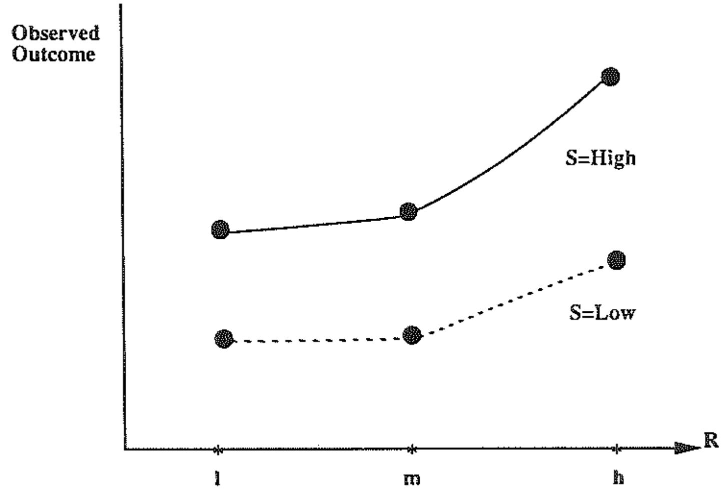

Fig. 3.2 Mean outcomes in a hypothetical factorial experiment.

Figure 3.2 uses the 3×2 example to illustrate direct and interactive effects. The vertical axis is the observed outcome, say market efficiency. The first treatment variable *R,* say elasticity of demand and supply, appears on the horizontal axis and the second variable *S,* say payoff intensity, shows up in the two curves labeled S = High and S = Low. The curves themselves connect the hypothetical mean outcomes in each treatment. The distance between the curves measures the direct effect of variable *S.* When the curves are parallel, there is no interaction between *R* and *S,* but when the gap between the curves widens as in Figure 3.1, there is a positive *RS* interaction. Chapter 7 will discuss the issue more extensively.

The factorial design is a bit less robust than the fully randomized design because experimenter errors in assigning treatments and missing trials (from computer glitches or no-show subjects, for instance) more seriously impair the data analysis. Indeed, if these problems are frequent, the factorial design becomes indistinguishable from the completely randomized.

Another problem with the basic factorial design is that the number of required trials increases quickly as the number of factors increases. Suppose, for example, you chose only two levels for each treatment variable. Even then, you need $2^4 = 16$ trials for 4 factors and $2^8 = 256$ trials for eight factors to run just a single replication! The problem is serious because there are many potentially important nuisance variables in some economic environments.

The fractional factorial design alleviates the problem. The basic idea is to run a balanced subset of the factorial design. To take the simplest example, suppose you have three variables, each with two levels denoted + and −, and can conduct only four trials. That is, you can run only half of the eight possible treatments (+ + +, + + −, + − +, + − −, − + +, − + −, − − +, and − − −). Your first thought might be just to run the first four treatments on the list, or every other treatment, but a moment's reflection shows that these choices are unbalanced because some variables are held constant or some pairs of variables are correlated. You get a balanced subset of treatments if you impose the restriction that the third sign is the product of the first two. Then the subset of treatments you run is + + +, + − −, − + −, and − − +. If you run this subset (in random order, of course!), then you have a half factorial 2×2×2 design. If you are a geometric thinker, you can visualize the balance of this design by thinking of each possible treatment combination as a corner of the unit cube in the space of the three treatment variables. For example, + + − could label the upper left back corner and − − + label the lower right front corner. The chosen treatments' center of mass is the center of the cube, and the center of mass on each face is the center of the face. Each level of each treatment variable appears in the same number of trials (2) and each pair of treatment variables is orthogonal.

Conceptually (although not visually) it is straightforward to generalize to more treatment variables and to smaller replication fractions. For example, Copeland and Friedman (1987) use a half-factorial 2×2×2×2 design in an asset-market experiment, where the fourth treatment variable (infocontent, a focus variable that defines the informational complexity of the environment) is constrained to be the product of the first three treatment variables (two nuisance variables called learnops and paymethod and another focus variable called infoarrival). A more dramatic example is given by Box et al. (1978, p. 394). They present a 2^7 sixteenth-factorial design for determining which of seven variables (seat position, handlebar position, tire pressure, etc.) affect a bicyclist's performance. Only 8 trials are required, compared to 128 in the full once-replicated factorial.

The elegance and economy of the fractional factorial design come at a price. The design obviously is less robust than a randomized design; it loses appeal if you are not confident of your ability to conduct all trials flawlessly. (If you are confident, the design has a subtle advantage: You can complete the factorial design if it turns out you can run additional trials.) The other disadvantage is inherent in the design. The fractional factorial achieves balance in a subset of the possible treatments by systematically confounding some direct effects with some interactions. The simple half-factorial 2×2×2 example confounds the third variable with the pairwise interaction of the first two variables, for instance. This disadvantage is not always serious. If you know that some pairwise or higher-order interactions are negligible, then you can harmlessly confound them.

We close this section with some background information for readers who wish to learn more about classical experimental design. R. A. Fisher and his colleagues developed most of the concepts presented in this chapter between 1910 and 1940. Much of the terminology comes from agricultural experiments; blocks, for example, originally referred to adjacent rectangular pieces of land, and a split-plot design (a type of randomized block) originally involved subdividing such a block for one treatment variable.

Statisticians with a combinatorial bent noticed that further efficiency gains theoretically arise from imposing additional symmetries on block and factorial designs. For instance, in testing four tire brands (*a*, *b*, *c*, and *d*) using four test cars, you could require not only the ordinary blocking condition that each car uses each brand, but also balance the assignment of tires to the four wheels of the test cars – say, use the order *abcd* for the four wheels in the first car, *dabc* in the second, *cdab* in the third, and *bcda* in the fourth car. This design is called Latin square after its diagrammatic representation, and it has higher-dimensional analogues called Graeco-Latin and hyper-Graeco-Latin designs. Such constructions quickly become quite Baroque and are not at all robust to missing trials and so forth.

The interested reader can find dozens of advanced books on experimental design, mostly of the 1950–70 vintage, in the QA279 section (under the Library of Congress system) and other sections of a good library. In writing this chapter we relied most heavily on Box et al., (1978) as well as Campbell and Stanley (1966), and Kirk (1982).

3.5 **Practical advice**

Theoretical considerations regarding experimental design do have practical consequences. Drawing on the theory, we offer some general advice regarding typical nuisance variables, the choice of constant and treatment variables, and the general conduct of experiments.

3.5.1 Chronic nuisances

Remember that the distinction between nuisance and focus variables depends on your purpose. Experience and learning, for example, serious because there are many potentially important nuisance variables in some economic environments.

The fractional factorial design alleviates the problem. The basic idea is to run a balanced subset of the factorial design. To take the simplest example, suppose you have three variables, each with two levels denoted $+$ and $-$, and can conduct only four trials. That is, you can run only half of the eight possible treatments ($+++$, $++-$, $+-+$, $+--$, $-++$, $-+-$, $--+$, and $---$). Your first thought might be just to run the first four treatments on the list, or every other treatment, but a moment's reflection shows that these choices are unbalanced because some variables are held constant or some pairs of variables are correlated. You get a balanced subset of treatments if you impose the restriction that the third sign is the product of the first two. Then the subset of treatments you run is $+++$, $+--$, $-+-$, and $--+$. If you run this subset (in random order, of course!), then you have a half factorial $2\times2\times2$ design. If you are a geometric thinker, you can visualize the balance of this design by thinking of each possible treatment combination as a corner of the unit cube in the space of the three treatment variables. For example, $++-$ could label the upper left back corner and $--+$ label the lower right front corner. The chosen treatments' center of mass is the center of the cube, and the center of mass on each face is the center of the face. Each level of each treatment variable appears in the same number of trials (2) and each pair of treatment variables is orthogonal.

Conceptually (although not visually) it is straightforward to generalize to more treatment variables and to smaller replication fractions. For example, Copeland and Friedman (1987) use a half-factorial $2\times2\times2\times2$ design in an asset-market experiment, where the fourth treatment variable (infocontent, a focus variable that defines the informational complexity of the environment) is constrained to be the product of the first three treatment variables (two nuisance variables called learnops and paymethod and another focus variable called infoarrival). A more dramatic example is given by Box et al. (1978, p. 394). They present a 2^7 sixteenth-factorial design for determining which of seven variables (seat position, handlebar position, tire pressure, etc.) affect a bicyclist's performance. Only 8 trials are required, compared to 128 in the full once-replicated factorial.

The elegance and economy of the fractional factorial design come at a price. The design obviously is less robust than a randomized design; it loses appeal if you are not confident of your ability to conduct all trials flawlessly. (If you are confident, the design has a subtle advantage: You can complete the factorial design if it turns out you can run additional trials.) The other disadvantage is inherent in the design. The fractional factorial achieves balance in a subset of the possible treatments by systematically confounding some direct effects with some interactions. The simple half-factorial $2\times2\times2$ example confounds the third variable with the pairwise interaction of the first two variables, for instance. This disadvantage is not always serious. If you know that some pairwise or higher-order interactions are negligible, then you can harmlessly confound them.

We close this section with some background information for readers who wish to learn more about classical experimental design. R. A. Fisher and his colleagues developed most of the concepts presented in this chapter between 1910 and 1940. Much of the terminology comes from agricultural experiments; blocks, for example, originally referred to adjacent rectangular pieces of land, and a split-plot design (a type of randomized block) originally involved subdividing such a block for one treatment variable.

Statisticians with a combinatorial bent noticed that further efficiency gains theoretically arise from imposing additional symmetries on block and factorial designs. For instance, in testing four tire brands (*a, b, c,* and *d*) using four test cars, you could require not only the ordinary blocking condition that each car uses each brand, but also balance the assignment of tires to the four wheels of the test cars – say, use the order *abcd* for the four wheels in the first car, *dabc* in the second, *cdab* in the third, and *bcda* in the fourth car. This design is called Latin square after its diagrammatic representation, and it has higher-dimensional analogues called Graeco-Latin and hyper-Graeco-Latin designs. Such constructions quickly become quite Baroque and are not at all robust to missing trials and so forth.

The interested reader can find dozens of advanced books on experimental design, mostly of the 1950–70 vintage, in the QA279 section (under the Library of Congress system) and other sections of a good library. In writing this chapter we relied most heavily on Box et al., (1978) as well as Campbell and Stanley (1966), and Kirk (1982).

3.5 **Practical advice**

Theoretical considerations regarding experimental design do have practical consequences. Drawing on the theory, we offer some general advice regarding typical nuisance variables, the choice of constant and treatment variables, and the general conduct of experiments.

3.5.1 Chronic nuisances

Remember that the distinction between nuisance and focus variables depends on your purpose. Experience and learning, for example, are nuisances if you want to test a static theory but are focus variables if you want to characterize behavioral change over time. This chapter has already mentioned most of the important nuisance variables you typically face in conducting an economics experiment, and suggested ways for dealing with them. Chapters 4 and 7 provide a more systematic discussion, but a quick summary may be useful at this point.

1. Experience and learning: Subjects' behavior changes over time as they come to better understand the laboratory environment. When this is a nuisance, control it as a constant by using only experienced subjects, or control it as a treatment (blocking variable) by using a balanced switchover design.
2. Noninstitutional interactions: Subjects' behavior may be affected by interactions outside the laboratory institution. For example, sellers may get together during a break and agree to maintain high prices. Careful monitoring during the break, or a change in parameters after the break, therefore may be advisable.
3. Fatigue and boredom: Subjects' behavior may change over time simply as a result of boredom or fatigue. For example, after playing strategy A for 58 periods in a repeated prisoner's dilemma, a subject may choose strategy B (defect) just to relieve the tedium. We recommend occasional payoff switchovers and planned sessions of at most two hours for most experiments.
4. Selection biases: The subjects or their behavior may be unrepresentative because their selection was biased. For example, self-selection may upwardly bias self-reported sexual activity when only the most talkative choose to respond to your questionnaire. Experimenter selection may be biased when students in an advanced finance class are recruited for an asset-market experiment. Recognizing the problem is the key step in finding ways to deal with selection biases.
5. Subject or group idiosyncrasies: A subject's background or temperament may lead to unrepresentative behavior. A group of subjects somehow may reinforce each other in unusual behavior patterns. Replication with different subjects therefore is essential.

3.5.2 Disposition of variables

We offer the following suggestions on choosing treatment and constant variables.

1. Control all controllable variables. Otherwise your data will be less informative than they could be.
2. Control focus variables as treatments. Use widely separated levels to sharpen the contrasts. Use two levels and skip intermediate levels unless you are interested in possibly nonlinear effects.
3. When you suspect that a nuisance variable interacts with a focus variable, consider controlling the nuisance as a treatment. Two levels often suffice.
4. Control most nuisances as constants to keep down complexity and cost. Even a nuisance with large effects can harmlessly be held constant as long as its effects are independent of the focus variables' effects.
5. Vary your treatments independently to maximize the resolution power of your data and to avoid confounding.

3.5.3 Phases of experimentation

A laboratory investigation typically proceeds in phases. The preliminary phase identifies the specific issues to be investigated and the essential aspects of the laboratory environment. The next phase consists of one or more pilot experiments. Here you complete the specification of the laboratory environment, prepare instructions for subjects, and conduct the pilot experiments, perhaps with unpaid subjects at first. The results usually lead to improving (simplifying) the instructions and the environment. At this point you should choose the focus and important nuisance variables you will use as treatments; the suggestions in the previous subsection may help.

Now you are ready to begin the formal part of your research by conducting a set of exploratory experiments. You should pick a simple design capable of detecting gross effects of the treatment variables, perhaps a fractional factorial or a $k = 1$ factorial. When you analyze the data you may decide to hold constant some variables that seem to have no interesting effects or interactions. Possibly you will want to adjust the environment or introduce a new treatment variable on the basis of the exploratory data. If you are exploring a new area, you may well discover at this point that major changes in instructions or treatments are necessary. If so, you will probably relabel your work so far as preliminary, and try the second phase again.

The final phase consists of follow-up experiments intended to provide definitive evidence on your chosen issues. Try to reserve 50 to 75 percent of your budget for this phase. If the results of the exploratory experiments seem clear-cut, you may choose simply to replicate them in the are nuisances if you want to test a static theory but are focus variables if you want to characterize behavioral change over time. This chapter has already mentioned most of the important nuisance variables you typically face in conducting an economics experiment, and suggested ways for dealing with them. Chapters 4 and 7 provide a more systematic discussion, but a quick summary may be useful at this point.

1. Experience and learning: Subjects' behavior changes over time as they come to better understand the laboratory environment. When this is a nuisance, control it as a constant by using only experienced subjects, or control it as a treatment (blocking variable) by using a balanced switchover design.
2. Noninstitutional interactions: Subjects' behavior may be affected by interactions outside the laboratory institution. For example, sellers may get together during a break and agree to maintain high prices. Careful monitoring during the break, or a change in parameters after the break, therefore may be advisable.
3. Fatigue and boredom: Subjects' behavior may change over time simply as a result of boredom or fatigue. For example, after playing strategy A for 58 periods in a repeated prisoner's dilemma, a subject may choose strategy B (defect) just to relieve the tedium. We recommend occasional payoff switchovers and planned sessions of at most two hours for most experiments.
4. Selection biases: The subjects or their behavior may be unrepresentative because their selection was biased. For example, self-selection may upwardly bias self-reported sexual activity when only the most talkative choose to respond to your questionnaire. Experimenter selection may be biased when students in an advanced finance class are recruited for an asset-market experiment. Recognizing the problem is the key step in finding ways to deal with selection biases.
5. Subject or group idiosyncrasies: A subject's background or temperament may lead to unrepresentative behavior. A group of subjects somehow may reinforce each other in unusual behavior patterns. Replication with different subjects therefore is essential.

3.5.2 Disposition of variables

We offer the following suggestions on choosing treatment and constant variables.

1. Control all controllable variables. Otherwise your data will be less informative than they could be.
2. Control focus variables as treatments. Use widely separated levels to sharpen the contrasts. Use two levels and skip intermediate levels unless you are interested in possibly nonlinear effects.
3. When you suspect that a nuisance variable interacts with a focus variable, consider controlling the nuisance as a treatment. Two levels often suffice.
4. Control most nuisances as constants to keep down complexity and cost. Even a nuisance with large effects can harmlessly be held constant as long as its effects are independent of the focus variables' effects.
5. Vary your treatments independently to maximize the resolution power of your data and to avoid confounding.

3.5.3 Phases of experimentation

A laboratory investigation typically proceeds in phases. The preliminary phase identifies the specific issues to be investigated and the essential aspects of the laboratory environment. The next phase consists of one or more pilot experiments. Here you complete the specification of the laboratory environment, prepare instructions for subjects, and conduct the pilot experiments, perhaps with unpaid subjects at first. The results usually lead to improving (simplifying) the instructions and the environment. At this point you should choose the focus and important nuisance variables you will use as treatments; the suggestions in the previous subsection may help.

Now you are ready to begin the formal part of your research by conducting a set of exploratory experiments. You should pick a simple design capable of detecting gross effects of the treatment variables, perhaps a fractional factorial or a $k = 1$ factorial. When you analyze the data you may decide to hold constant some variables that seem to have no interesting effects or interactions. Possibly you will want to adjust the environment or introduce a new treatment variable on the basis of the exploratory data. If you are exploring a new area, you may well discover at this point that major changes in instructions or treatments are necessary. If so, you will probably relabel your work so far as preliminary, and try the second phase again.

The final phase consists of follow-up experiments intended to provide definitive evidence on your chosen issues. Try to reserve 50 to 75 percent of your budget for this phase. If the results of the exploratory experiments seem clear-cut, you may choose simply to replicate them in the follow-up phase. If the exploratory experiments suggest subtle but relevant direct effects or interactions among your variables, you may choose a more elaborate design.

A final piece of advice. Don't get too fancy in designing your experiments, especially in your first project. Begin with a proven design from related previous research by other authors, or use a simple version of one of the designs we have presented.

3.6 **Application: New market institutions**

We live in an era of rapid change in economic institutions. Existing markets have expanded and changed, and new markets have opened, in response to advances in computer and telecommunications technology and in response to political developments in Asia, and in Eastern as well as Western Europe. Even in the relatively stable markets of the United States, scandals and technological developments have spurred efforts to reform the primary market for U.S. government securities and the commodity exchanges.

How do we evaluate alternative market institutions? What kinds of market institutions will best promote efficient exchange in the new environments around the world? Existing economic theory and historical experience provide precious little guidance. Field experiments can be costly, as well as politically risky. Laboratory experiments can conveniently serve as test beds for new market institutions. New institutions can be tried out and refined in the laboratory before they are further tested and implemented in the field. This section discusses some of the test-bed work done so far and uses it to illustrate some of the basic principles and issues in experimental design.

> Laboratory experimentation can facilitate the interplay between the evaluation and modification of proposed new exchange institutions before field implementation. . . . Laboratory experiments allow one to investigate the incentive and performance properties of alternative exchange institutions, and, with respect to institutional design, they provide a low-cost means of trying, failing, altering, trying, etc. This process uses theory, loose conjecture, intuitions about procedural matters and, most important, repeat testing to understand and improve the features of the institutional rules being examined. (McCabe, Rassenti, and Smith, 1993, p. 309)

Two kinds of work are discernible in test-bed research. When the institutions are reasonably well-specified, an experiment can be designed using classical approaches discussed in this chapter in order to measure and compare their performance characteristics. The studies by Hong and Plott (1982) and by Grether and Plott (1984) described below fall into this *performance testing* branch of test-bed research. On the other hand, when the institution itself has to be designed through an iterative design-test-revise process, classical experimental design techniques usually cannot be applied to the overall process, although they may be useful for some phases of the project. This second branch, *developmental testing* is exemplified in Grether, Isaac, and Plott (1981), Plott and Porter (1989), the McCabe et al. (1993) effort to develop a uniform-price double auction, and the McCabe et al. (1988) effort to develop a "smart" market for natural gas. We shall now briefly touch on both branches of test-bed research.

3.6.1 Performance testing

Grether and Plott (1984) conducted some early test-bed experiments dealing with a controversy about existing market institutions. In May 1979 the U.S. Federal Trade Commission filed an antitrust suit against the four domestic producers of a gasoline additive, tetraethyl lead. The suit claimed that uncompetitive high prices were sustained by three institutional practices: advanced notification of price changes (AN), "most favored nation" guarantees to customers that nobody else will get a lower price (MFN), and "delivered pricing" quotes that include transportation cost (DP). The four lead producers argued that the institutional practices were a convenience to customers and had no anticompetitive effects.

In their laboratory study, Grether and Plott break the AN institution down into three focus variables: price publication with three levels (N = no seller publishes prices, L = the two largest sellers publish, and A = all sellers publish prices), price access with two levels (B = only buyers see published prices, and A = all buyers and all sellers see published prices), and advanced notice per se with two levels (Y = yes, a seller can change price only if it is announced in the previous period, and N = no advanced notice required). They made MFN a single two-level (Y or N) variable and omitted DP from their study. Even so, there are potentially $3 \times 2 \times 2 \times 2 = 24$ institutional treatments (i.e, conjunctions of the four treatment variables).

In order to keep the study within budget, Grether and Plott held constant most other relevant variables including supply–demand parameters (at a level chosen to resemble the field conditions) and the basic exchange institution (bilateral search using telephones). Some conjunctions of treatments are vacuous or uninteresting (e.g., access to prices when no sellers publish prices) and some are especially interesting (e.g.,