

### HOW TO DO IT 4.3

#### Describing a Posttest-Only Control Group Design

As the most used and recommended of all the Campbell and Stanley true experiment types, the posttest-only control group design is likely one you will use often. Unlike experiments that use the Solomon four-group design, most studies that use this ubiquitous design do not announce themselves formally. Rather, you can determine they use this typology by looking for mention of a control group and random assignment, and no mention of a pretest.

Here are two examples; in the first, it does specify it is a posttest-only design, followed by the second, which does not.

**Coleman, Renita, Paul Lieber, Andrew Mendelson, and David Kurpius. 2008. "Public Life and the Internet: If You Build a Better Website, Will Citizens Become Engaged?" *New Media & Society* 10 (2): 179–201.**

"This study used a post-test, control group experimental design. The experimental stimulus for this study was a website on the topic of the state budget created by mass communication students in a class in website development . . ." (p. 188). "The control group website was the official state government website on the state budget. It was created without usability tests or knowledge of any of the issues described above, which guided the creation of the experimental website. It was important that the control site was on the same topic as the experimental site in order to rule out the possibility of effects due to the subject matter rather than content, appearance or navigation . . ." (p. 189). "The 60 participants were randomly assigned to view either the control website or the experimental website" (p. 190).

**Bennion, Elizabeth A., and David W. Nickerson. 2011. "The Cost of Convenience: An Experiment Showing E-Mail Outreach Decreases Voter Registration." *Political Research Quarterly* 64 (4): 858–869.**

"The design of the e-mail experiment itself was straightforward. Students were randomly assigned to one of three conditions: (1) a control group receiving no e-mail, (2) a treatment group receiving three e-mails from an administrator such as the university president or dean of students, or (3) a treatment group receiving three e-mails from a student leader—usually the student body president. The e-mails were brief, explaining why registration is important and providing a link to the Rock the Vote online registration tool" (p. 862).

independent—that is, some of the same people may be in more than one group—and, finally, that quasi experiments cannot control for all the extraneous factors that true experiments do. These features of random assignment, independence, and control are not always possible. When that is the case, a quasi experiment is considered a viable option, and even preferable in some cases. Some things simply cannot be "assigned." For example, researchers cannot ethically assign someone to smoke and someone else to not smoke in order to have randomly assigned treatment and control groups. Experiments in business and education settings especially make it difficult to have complete control over the research.<sup>28</sup> Businesses involved in a study may want to handpick participants for



iStock.com/skymaster

When students in intact classrooms are used in experiments, these are known as *quasi experiments* because the subjects are not randomly assigned.

some reason, or want everyone to be treated the same and not appear to be showing favoritism. In other cases, participants unwittingly self-select the group they are in. For example, education research frequently works with intact classrooms rather than classes whose students were randomly assigned. Educators may assign students to classes because they want diversity or a mix of boys and girls, do not want twins in the same class, or want to keep two children together. At the college level, it is obvious that there is something individually different about people who sign up for a class in experimental design than those who register for ethnographic methods. Students might choose one course over another because they heard a certain teacher is good. Or it might be a timing issue—some students do not like getting up early in the morning, have part-time jobs on Wednesday afternoons, or want to leave town on Fridays. Whatever is at work to create these individual differences could compromise the validity of an experiment.

### For Review-No Commercial Use(2023)

Only random assignment is an antidote to such individual differences affecting results. Research can still uncover important knowledge from quasi experiments, but they bring different threats to validity than do true experiments, and these must be addressed and documented. It is important to say in the final paper that subjects could not be randomly assigned and to give the reasons why. The paper also should explain what was done to minimize any issues with systematic differences—for example, were key variables measured and then used as covariates? Were statistical tests run to see if groups were equivalent on important variables? It is important to make a credible claim for being able to infer causality, or to generalize beyond the one case studied. Finally, quasi experiments cannot control all the extraneous factors that may cause an outcome the way true experiments can. Without random assignment, a quasi experiment cannot eliminate the possibility of uncontrolled variables confounding the results and impairing the ability to make causal claims. However, researchers should always attempt to measure possible confounding variables, then control for them statistically by using covariates. Every study has its weaknesses, including true experiments (see more about these in chapter 5). Limitations do not necessarily invalidate the importance of the findings, especially if this is the only way to study a particular phenomenon. Quasi experiments can feel less artificial than true experiments and are usually easier to conduct longitudinally than controlled experiments—all benefits. It is better to know what the study found than to not know anything, but readers should always be told about the limitations.

Triangulation, or confirming what has been found using one method with another, is also important. If different methods confirm the same finding—for example, if what is found in a true experiment in an artificial setting dovetails with what happens in the real world of a quasi experiment—then we can have more confidence that the findings represent something real.

One example of best practices in a quasi experiment in social work education used two similar classes to study whether online or face-to-face teaching would result in greater learning.<sup>29</sup> In this quasi experiment, the researchers used intact classrooms and so were unable to randomly assign students to classes with different teaching styles. To attempt to control as much as possible for individual student differences, they used as covariates the students' ages and grade point averages adjusted for grade inflation. They say some of the limitations are the unaccounted for extraneous variables and the lack of generalizability.

In a business study, the researcher examined how formal mentoring affected workers' ability to build interpersonal networks.<sup>30</sup> Those who were mentored represented the treatment group; those who were not mentored were the control group. In that study, participants were not randomly assigned; rather, the management of the company chose the employees who would be mentored based on their perceived potential for advancement within the organization. The researchers used a matched pairs design to control for confounding variables. One group was chosen specifically because they were perceived as superior to the other, but the company would not agree to random assignment. To decrease the uncertainty, the researcher used a "matched pairs" design, where every worker in the treatment group was paired with a worker in the control group who was as similar as possible on important characteristics—they had the same salary, performance rating in the prior year, length of time with the company, and were in the same office. The author of that study provides statistical evidence that matching was successful, showing that there were no significant differences between the treatment and control groups on other characteristics, including age, education, and the number of networking contacts each group made before mentoring began. More network contacts was one of the outcome goals of the program, so this assured that people in the treatment group were not more likely to network to begin with. Still, the workers in the treatment and control groups could have differed in other unknown ways that could have caused differences in outcomes that were supposed to be attributed to the mentoring treatment, so various other strategies were used to help overcome the lack of random assignment. However, because we can never think of all the other confounding variables, we can never be assured these were as successful as simple random assignment; more uncertainty always remains.

Another example of a quasi experiment is the study of an intervention where investigators interviewed abused and neglected children about their mental health and quality of life.<sup>31</sup>

This type of research typically only interviewed adults, as it was thought children were unreliable sources and did not need to be upset by these kinds of questions. In this study, the cases where only adults were interviewed represented the control group; the cases where both children and adults were interviewed represented the intervention. Random assignment to treatment or control group was not possible because the Dutch Medical Ethics Committee refused to approve it out of concern that the adult-only interview group might receive inferior care. The researchers were able to have treatment and control groups, but the participants could not be randomly assigned to them. The article notes the possibility of selection bias.

Beyond the pre-experimental, true, and quasi experimental designs in Campbell and Stanley's typology, researchers also employ natural or field experiments. These are briefly covered next, and there are many good books that treat these topics in depth (see the "Suggested Readings" section). Instead, this is offered as an introduction as food for thought about whether your topic is best suited for a true experiment, a quasi experiment, a natural or a field experiment.

## NATURAL EXPERIMENTS

For Review-No Commercial Use(2023)

Natural experiments are a subset of quasi experiments in that researchers do not randomly assign subjects to conditions or create the manipulation. Instead, natural experiments take advantage of some naturally occurring phenomenon that creates treatment and control groups. One group of people was exposed to something, and another group was not, in a natural setting. This is seen as approximating randomization as far as possible, what some researchers call *near random* or *as-if random*.<sup>32</sup> Nature or society creates the treatment or exposure, and researchers discover it after it has already occurred, then conceive of it as an experiment. Crasnow says it is really more of an observational study.<sup>33</sup> Because researchers did not design the treatment or intervention, they cannot control all other possible factors that could have caused the observed outcome. This could lead to plausible alternative explanations, which need to be explained and accounted for as much as possible.

Evaluations of programs designed to improve some condition in society are frequently conducted as natural experiments. One example is a study evaluating a program developed by the Maryland Network Against Domestic Violence that provided social service advocates for victims.<sup>34</sup> In this study, the police and social service organizations designed their own intervention—providing advocacy, safety planning, and referral services to women. The researchers came in afterward to study the efficacy of it.

At other times, researchers simply notice an “intervention” that takes place naturally and study it. For example, researchers compared criminal offenders who moved against those who did not to see if the locations of their crimes changed to be closer to their new homes.<sup>35</sup> The researchers did not design the intervention—assigning people to move—it just happened naturally. The researchers used a host of covariates to help keep the plausible alternative explanations to a minimum, including type of crime, time between crimes, where previous crimes were committed, and where previous homes were. Despite these precautions, a number of other plausible explanations are discussed, as they should be.

This is a good place to note that researchers sometimes use terms such as quasi experiment and natural experiment interchangeably; for precision of language, this book uses *quasi experiment* for studies where researchers design the intervention and *natural experiment* for studies that take advantage of some naturally occurring intervention. Neither of these uses random assignment. Articles that report true experiments or **laboratory experiments** with random assignment do not typically use the terms *true* or *laboratory* but are just called experiments, and refer to anything not conducted in its natural setting that uses random assignment.

Another example of a natural experiment, a classic in the field of communication, is the study of three Canadian towns: one that had no TV, one that had one TV channel, and another that had four channels. This was back in the days when television signals were broadcast over the airwaves through antennas and small boxes were the last to receive television. All three towns in the study were revisited three years after the no-TV town had gotten TV. This research added longitudinal knowledge about the short- and long-term nature of television’s effects on children’s aggressive behavior, reading skills, cognitive development, leisure activities, use of other media, sex-role attitudes, and other personality traits and attitudes.<sup>36</sup> In these experiments, people were assigned to conditions by forces other than researchers—that is, whatever led them to live in those towns. There is some treatment or exposure to something; in this example, it was exposure to television in varying levels—no TV, one TV channel, four channels—that leads to the ability to say that TV exposure caused any differences in outcomes. Because it was not a randomly assigned, in-laboratory experiment, there are many other things that could have caused the differences or confounded the outcomes,<sup>37</sup> but the researchers did everything they could think of to control for these.

Natural experiments are becoming increasingly popular in many different disciplines.<sup>38</sup> Researchers can take advantage of changes in the world to conduct natural experiments. For example, when business scholars studied the changes in Peru’s soft-drink market before and after the country entered into a free trade agreement with the United States, they used as a control group the country of Bolivia, which has no such agreement.<sup>39</sup> They selected the control country by matching important characteristics such as demographics, income growth,

population, and economic trade indicators. While the comparison countries are not exact, they are as close as the researchers could get. The authors discuss the limitations and include suggestions for future studies that would include more controls for wage rates, economic and political stability, and using more than one country as a comparison.

Another study in London looked at public health before and after the introduction of free bus rides for young people.<sup>40</sup> Potential confounds included the fact that other policies designed to change people's choice of transportation had also been introduced recently (e.g., higher charges for driving during congested times), a change in cultural attitudes in general (e.g., walking to prevent obesity increased), and that there was no control group (all people under eighteen were given free bus passes). They list other limitations of a natural experiment as including a weaker ability to make causal claims than true experiments and difficulty generalizing beyond the single case being studied. The paper explains well the trade-offs between a realistic setting and internal validity, and offers some solutions that include mixing designs and data collection. They say the results are “good enough” evidence and “as robust an evaluation as possible.”<sup>41</sup>

Advance planning allows researchers to obtain approval to study human subjects from their institutional review boards (IRBs) in advance. This is not possible in all situations that are ripe for natural experiments come with advance warning, however—for example, the study of how re-incarceration rates changed after Hurricane Katrina, forcing parolees to spread out around the state rather than be concentrated in particular neighborhoods,<sup>42</sup> or how attitudes toward welfare recipients changed before and after riots.<sup>43</sup> These studies can only be done if researchers use data collected before the event, receive approval after the fact, or are in the enviable position of having already obtained IRB approval for a study on a similar topic before the event happens and can quickly get approval for an amendment. Some even make a distinction between natural experiments and “nature's experiments,” but this book will not go into that.<sup>44</sup> It is never a bad thing to keep an eye open for opportunities such as this.

## FIELD EXPERIMENTS

---

While natural experiments typically do not have the benefit of random assignment because people are exposed to the experimental or control conditions based on natural factors outside the control of researchers, there is another type of experiment, the field experiment, that takes advantage of real-world settings but also employs random assignment to treatment and control groups.<sup>45</sup>

The term *field experiment* makes the distinction between experiments conducted in natural settings versus laboratory settings, even though not many “lab” experiments are conducted in actual laboratories anymore. For example, experiments that use survey software (covered in chapter 7) can be conducted in the subjects’ own homes. Field experimentalists consider any setting other than the environment under which something would naturally occur to be a lab. Anything conducted on a college campus is a lab experiment, unless the actual context is a college setting—for example, a study of cheating on tests. Another example would be having participants look at TV commercials on their home computer and evaluate them to be considered a lab experiment. Even though the subjects are in their own homes, they are not looking at the ads on TV and they know that they are doing so for a research study. That is another key distinction for field experimentalists; subjects should not be aware of being studied. Field experiments “strive to be as realistic and unobtrusive as possible.”<sup>46</sup> The focus on realism and subjects being unaware of being studied stems from concerns about misleading results due to people behaving differently when they are being watched, known as the Hawthorne effect, and reactivity, or participants wanting to give the “right answer” or the answer the experimenter wants. Whether these are serious problems with lab experiments is unclear, as few studies replicate experiments in both field and lab conditions in order to see if the results differ. Field experimentalists also note that what works in a lab setting might not work in the real world.<sup>48</sup> Some effects can be immediate and strong, so they show up in a lab experiment but decay over time and would show up weaker in a field experiment.

Another objection field experimentalists have to lab experiments is the potential lack of realism in the messages or stimuli created by researchers. These are all important concerns, and lab experimentalists should be careful to see that their stimuli are as realistic as possible, having practitioners in their field create or review the interventions. This is a topic of chapter 9 on creating stimuli.

Field experiments should be authentic on four dimensions: the participants, treatment, context, and outcome measures.<sup>49</sup> Participants should be real voters, not students pretending to be, for example. The treatment should be a real political debate, not one fielded by actors. The context should have voters watching TV in their living rooms, not with a group of strangers in a university classroom. And the outcome measures should be their votes or donations to the candidate, not self-reports of their intentions to vote or donate.

Field experiments can be more expensive and difficult than traditional lab experiments, and also ethically challenged given that one hallmark is that subjects do not know they are participating in an experiment (see chapter 11 on ethics). Gerber and Green discuss



in detail three of the most common challenges:<sup>50</sup> Briefly, noncompliance is when subjects that were assigned to one treatment actually got something else; attrition is when outcome measures are not obtained for every subject; and interference occurs when subjects talk to each other, compare notes, or remember treatments. These issues are minimized or nonexistent in the environment of a lab experiment.

Whereas lab experiments are commonly employed to test theoretical propositions, where tightly controlled conditions are important, field experiments are better for applied studies—for example, evaluations of a program's effectiveness.<sup>51</sup>

A few examples of field experiments:

To see if interacting directly with a politician could persuade people to change their attitudes about issues, their assessments of the politicians' qualities, and how they vote, researchers used online town hall meetings, noting that the only studies up to that point had used laboratory settings that simulated only a few of the characteristics of personal interactions.<sup>52</sup> Previous research was equivocal about whether real-life results were the same as those from hypothetical settings. So the researchers recruited U.S. senators and members of the U.S. House of Representatives to interact with their constituents in a real-time online forum. The intervention in this field experiment was created by the researchers but was more true-to-life than a mock town hall with an actor playing a politician. Citizens who participated were randomly assigned either to participate in the online discussion with the politician or to the control group, which only received reading material with background about the issues. Those in the treatment group got the same reading material but also participated in the online session. You can see some of the same kinds of effort exhibited here as Philip Zimbardo did in recruiting the Palo Alto police to “arrest” his prison experiment subjects.

Other researchers did a randomized field experiment in the Netherlands on the effectiveness of an extended day program on elementary students' math and language learning.<sup>53</sup> They noted that the research on program effects in education rarely used randomized experiments and that quasi experiments that did not have proper controls were the norm. In this study, the researchers randomly selected students and offered them the chance to participate in the program rather than allowing them or their teachers to select who participated. Like studies before them that used randomized experiments and found small to nonexistent effects, this one did not find any effects. It is important to know when different methods generate opposite results; for example, on this topic, quasi experiments without random assignment were likely to show effects, but more rigorous randomized true experiments were not. In this case,



the field experiment, using random assignment, helped educators know which results to have more confidence in.

As with the terms quasi experiment and natural experiment, researchers also use the term *field experiment* to mean different things. In this book, I advocate the use of the term field experiment to mean an experiment that is conducted in a natural or real-life setting and also uses random assignment. I use quasi and natural experiments when subjects are not randomly assigned to treatment and control conditions.

This book is devoted primarily to true experiments, also called lab experiments, although the concepts covered here are applicable to other types of experiments as well. Those conducting a quasi, natural, or field experiment should also consult texts that address the specific issues associated with those designs.

In summary, there is no perfect experiment. One cannot usually have the benefit of a real-world setting and also have random assignment. The best researchers can do to understand a particular phenomenon is conduct many different studies of varying designs that have at least one of these features, using different contexts. Replication under different conditions affords more confidence that what has been found is real. In fact, being bold enough to understand when an opportunity presents itself to study something by an unconventional method is one of the hallmarks of a creative researcher.

This chapter is not an exhaustive compilation of all the many different experimental designs available. Many other options exist, and the creative experimentalist will be open to discovering new and better ways to study important causes and effects. The next chapter begins this textbook's sole focus on true or laboratory experiments, starting with issues of internal and external validity.

### Common Mistakes

- Using pretests when they are not truly necessary
- Reporting an experiment as “exploratory” because it contains flaws
- Failing to randomly assign subjects to conditions. Studies that aspire to be true experiments but fail to use randomization are rarely published.

## Test Your Knowledge

1. From Campbell and Stanley's typology of experimental designs, which one is considered the gold standard?
  - a. Pretest–posttest control group design
  - b. Solomon four-group design
  - c. Posttest-only control group design
  - d. Static group comparison design
2. From Campbell and Stanley's typology of experimental designs, which one is most recommended today?
  - a. Pretest–posttest control group design
  - b. Solomon four-group design
  - c. Posttest-only control group design
  - d. Static group comparison design
3. A pretest is essential for an experiment to be considered a true experiment.
  - a. True
  - b. False
4. What is the *main* reason for using random assignment?
  - a. To make sure subjects in treatment and control groups are equivalent on important characteristics.
  - b. To make sure the same number of subjects are in the treatment and control groups.
  - c. To ensure that the same people are not in more than one group.
  - d. To make it fair to all subjects.
5. Which type of experiment uses random assignment in a naturally occurring setting?
  - a. Lab experiment
  - b. Quasi experiment
  - c. Natural experiment
  - d. Field experiment
6. Which of the following designs is recommended when a pretest would NOT threaten to change subjects' performance?
  - a. Static group comparison
  - b. Pretest–posttest control group

For Review-No Commercial Use(2023)

- c. One-group pretest–posttest design
  - d. Quasi experiment
7. Which of the following is a strength of designs with pretests?
- a. It gives a baseline measure to compare any changes against.
  - b. Being observed or measured twice may cause changes in the subjects' performance.
  - c. Pretests are essential to true experimental designs.
  - d. It ensures the groups are equal on important characteristics that could otherwise cause any changes.
8. A study used two similar intact classes to see whether online or face-to-face teaching would result in greater learning. This is an example of:
- a. A natural experiment
  - b. A field experiment
  - c. A laboratory experiment
  - d. A quasi experiment
9. An experiment that is conducted in a natural or real-life setting and also uses random assignment is: **For Review-No Commercial Use(2023)**
- a. A quasi experiment
  - b. A natural experiment
  - c. A field experiment
  - d. A true experiment
10. Which of the following allows the researcher to compare the differences before and after treatment, and also to tell if there were any effects of a pretest?
- a. Static group comparison
  - b. Solomon four-group design
  - c. Pretest–posttest control group
  - d. One-group pretest–posttest design

#### Answers

- |      |      |      |       |
|------|------|------|-------|
| 1. b | 4. a | 7. a | 9. c  |
| 2. c | 5. d | 8. d | 10. b |
| 3. b | 6. b |      |       |

## Application Exercises

1. Use Google Scholar.com or your school library's database to find studies that use experimental designs in your discipline. Read three of the experiments that interest you most and identify the design of the experiment; is it a true or laboratory experiment, quasi experiment, natural or field experiment? Is it a pretest–posttest control group design, Solomon four-group design, or something else? What are some of the limitations, and how did the authors address them?
2. Examine the literature about your topic for the methodologies used. A grid-type chart will help you see which methods have been most used. Of the twenty-five-plus articles you read for the literature review assignment in chapter 3, categorize them by method, including critical essay, focus group, interviews, ethnography, content analysis, survey, and experiment, among others. Which method has been most used? Is it appropriate for an experiment to be conducted now that establishes cause and effect?

## Suggested Readings

- Campbell, Donald T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Crasnow, S. 2015. "Natural Experiments and Pluralism in Political Science." *Philosophy of the Social Sciences* 45 (4/5): 424–441.
- Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: Norton.
- Levy, Y., T. J. Ellis, and T. Cohen. 2011. "A Guide for Novice Researchers on Experimental and Quasi-Experimental Studies in Information Systems Research." *Interdisciplinary Journal of Information, Knowledge and Management* 6: 151–161.

## Notes

1. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. (Chicago: Rand McNally, 1963).
2. Ibid., 6.
3. David W. Catron and Claudia C. Thompson, "Test-Retest Gains in WAIS Scores after Four Retest Intervals," *Journal of Clinical Psychology* 35, no. 2 (1979): 352–357.
4. Ibid.
5. Brendon R. Barnes, "The Hawthorne Effect in Community Trials in Developing Countries," *International Journal of Social Research Methodology* 13, no. 4 (2010): 357–370.
6. John G. Adair, "The Hawthorne Effect: A Reconsideration of the Methodological Artifact," *Journal of Applied Psychology* 69, no. 2 (1984): 334–345; Ryan Olson et al., "What We Teach Students About the Hawthorne Studies: A Review of Content Within

- a Sample of Introductory I-O and OB Textbooks,” *The Industrial Organization Psychologist* 41 (2004): 23–39.
7. E. Mayo, 1933; Chen-Bo Zhong and Julian House, “Hawthorne Revisited: Organizational Implications of the Physical Work Environment,” *Research in Organizational Behavior* 32 (2012): 3–22.
  8. Olson et al., “What We Teach Students.”
  9. Henry A. Landsberger, *Hawthorne Revisited. Management and the Worker: Its Critics, and Developments in Human Relations in Industry* (Ithaca, NY: Cornell University, 1958).
  10. J. G. Adair, D. Sharpe, and C. Huynh, “Hawthorne Control Procedures in Educational Experiments: A Reconsideration of Their Use and Effectiveness,” *Review of Educational Research* 59, no. 2 (1989): 215–227.
  11. Barnes, “The Hawthorne Effect in Community Trials”; Zhong and House, “Hawthorne Revisited.”
  12. Baptiste Leurent et al., “Monitoring Patient Care through Health Facility Exit Interviews: An Assessment of the Hawthorne Effect in a Trial of Adherence to Malaria Treatment Guidelines in Tanzania,” *BMC Infectious Diseases* 16 (2016): 1–9.
  13. Jim McCambridge, John Witton, and Diana R. Elbourne, “Systematic Review of the Hawthorne Effect: New Concepts Are Needed to Study Research Participation Effects,” *Journal of Clinical Epidemiology* 67, no. 3 (2014): 267–277; Magnus Hansson and Rune Wigblad, “Recontextualizing the Hawthorne Effect,” *Scandinavian Journal of Management* 22, no. 2 (2006): 120–137.
  14. Martin T. Orne, “Demand Characteristics and the Concept of Quasi Controls,” in *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow’s Classic Books*, ed. Robert Rosenthal and Ralph L. Rosnow (Oxford: Oxford University Press, 2009): 110–137; D. Steele-Johnson et al., “Goal Orientation and Task Demand Effects on Motivation, Affect, and Performance,” *Journal of Applied Psychology* 85, no. 5 (2000): 724–738.
  15. Martin T. Orne, “On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications,” *American Psychologist* 17 (1962): 776–783; “Demand Characteristics and the Concept of Quasi-Controls,” in *Artifact in Behavioral Research*, ed. R. Rosenthal and R. Rosnow (New York: Academic Press, 1969), 143–179.
  16. Barnes, “The Hawthorne Effect in Community Trials.”
  17. Adair, “The Hawthorne Effect.”
  18. David Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* (New York: W. H. Freeman, 2001).
  19. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  20. Kaanan Butor-Bhavsar, John Witton, and Diana Elbourne, “Can Research Assessments Themselves Cause Bias in Behaviour Change Trials? A Systematic Review of Evidence from Solomon 4-Group Studies,” *PLoS ONE* 6, no. 10 (2011): 1–9; Campbell and Stanley, *Experimental and Quasi-Experimental Designs*; Shlomo Sawilowsky and D. Lynn Kelley, “Meta-Analysis and the Solomon Four-Group Design,” *Journal of Experimental Education* 62, no. 4 (Summer 1994): 361.
  21. R. Barker Bausell, *Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist* (Thousand Oaks, CA: Sage, 1993), 10.
  22. Yair Levy, Timothy J. Ellis, and Eli Cohen, “A Guide for Novice Researchers on Experimental and Quasi-Experimental Studies in Information Systems Research,” *Interdisciplinary Journal of Information, Knowledge & Management* 6 (2011): 154.
  23. Stephen Gorard, “The Propagation of Errors in Experimental Data Analysis: A Comparison of Pre- and Post-Test Designs,” *International Journal of Research & Method in Education* 36, no. 4 (2013): 372–385.
  24. Ibid., 372.
  25. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  26. V. L. Willson and R. R. Putnam, “A Meta-Analysis of Pretest Sensitization Effects in Experimental Design,” *American Educational Research Journal* 19 (1982): 249–258.
  27. For an example of this, see Charles Boy Kromann, Morten L. Jensen, and Charlotte Ringsted, “The Effect of Testing on Skills Learning,” *Medical Education* 43, no. 1 (2009): 21–27.
  28. Levy, Ellis, and Cohen; J. W. Creswell, *Educational Research: Planning, Conducting, and Evaluating*

- Quantitative and Qualitative Research*, 2nd ed. (Upper Saddle River, NJ: Pearson, 2005).
29. Ralph Woehle and Andrew Quinn, "An Experiment Comparing HBSE Graduate Social Work Classes: Face-to-Face and at a Distance," *Journal of Teaching in Social Work* 29, no. 4 (2009): 418–430.
  30. Sameer B. Srivastava, "Network Intervention: Assessing the Effects of Formal Mentoring on Workplace Networks," *Social Forces* 94, no. 1 (September 2015): 427–452.
  31. Froukje Snoeren et al., "Design of a Quasi-Experiment on the Effectiveness and Cost-Effectiveness of Using the Child-Interview Intervention During the Investigation Following a Report of Child Abuse and/or Neglect," *BMC Public Health* 13, no. 1 (2013): 1–16.
  32. Sharon Crasnow, "Natural Experiments and Pluralism in Political Science," *Philosophy of the Social Sciences* 45, no. 4/5 (2015): 429.
  33. Ibid.
  34. Jill Theresa Messing et al., "The Oklahoma Lethality Assessment Study: A Quasi-Experimental Evaluation of the Lethality Assessment Program," *Social Service Review* 89, no. 3 (2015): 499–536.
  35. Andrew Wheeler, "The Moving Home Effect: A Quasi Experiment Assessing Effect of Home Location on the Offence Location," *Journal of Quantitative Criminology* 28, no. 4 (2012): 587–606.
  36. Tannis MacBeth Williams, *The Impact of Television: A Natural Experiment in Three Communities*, ed. Tannis MacBeth Williams (Orlando, FL: Academic Press, 1986).
  37. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  38. Crasnow, "Natural Experiments and Pluralism."
  39. Phillip Baker et al., "Trade and Investment Liberalization, Food Systems Change and Highly Processed Food Consumption: A Natural Experiment Contrasting the Soft-Drink Markets of Peru and Bolivia," *Globalization and Health* 12 (2016): 1–13.
  40. Judith Green et al., "Integrating Quasi-Experimental and Inductive Designs in Evaluation: A Case Study of the Impact of Free Bus Travel on Public Health," *Evaluation* 21, no. 4 (2015): 391–406.
  41. Ibid., 395–396.
  42. David S. Kirk, "A Natural Experiment of the Consequences of Concentrating Former Prisoners in the Same Neighborhoods," *Proceedings of the National Academy of Sciences of the United States of America* 112, no. 22 (2015): 6943–6948.
  43. Aaron Reeves and Robert de Vries, "Does Media Coverage Influence Public Attitudes Towards Welfare Recipients? The Impact of the 2011 English Riots," *British Journal of Sociology* 67, no. 2 (2016): 281–306.
  44. Mary S. Morgan, "Nature's Experiments and Natural Experiments in the Social Sciences," *Philosophy of the Social Sciences* 43, no. 3 (2013): 341–357.
  45. Alan S. Gerber and Donald P. Green, *Field Experiments: Design, Analysis, and Interpretation* (New York: W. W. Norton, 2017).
  46. Ibid., 9.
  47. Ibid.
  48. Ibid.
  49. Ibid.
  50. Ibid.
  51. Ibid.
  52. William Minozzi et al., "Field Experiment Evidence of Substantive, Attributional, and Behavioral Persuasion by Members of Congress in Online Town Halls," *Proceedings of the National Academy of Sciences* 112, no. 13 (2015): 3937–3942.
  53. Erik Meyer and Chris Van Klaveren, "The Effectiveness of Extended Day Programs: Evidence from a Randomized Field Experiment in the Netherlands," *Economics of Education Review* 36 (2013): 1–11.



# INTERNAL AND EXTERNAL VALIDITY

*No issue . . . has a longer half-life or recurs with more regularity than the argument over the validity and generalizability of findings obtained from social scientific experiments.<sup>1</sup>*

—John A. Courtright

For Review-No Commercial Use(2023)

## LEARNING OBJECTIVES

- Illustrate internal and external validity and the trade-offs.
- Identify how generalizability is achieved in experiments.
- Compare logical inference versus statistical inference.
- Discuss the role of replication in experimentation.
- Explain how random assignment provides internal validity.
- Identify the seven classes of extraneous variables that jeopardize internal validity.

The previous chapter looked at different kinds of experiments—lab or true, quasi, natural, and field—and concluded there is no perfect experiment. As in life, experiments involve trade-offs, and that is especially true of the two sides of **validity**—external and internal. Like children on a seesaw, one type of validity tends to go up as the other



goes down. In social science experiments, the goal is to maximize both, but that is usually difficult to achieve.<sup>2</sup> As the opening quote shows, this debate can continue for a very long time. To be “valid” means something is well founded, accurate, or authoritative. In the case of science, it is especially important that the conclusions of research be valid or true, at least with an acceptable level of probability. The topic of validity is especially important for experiments, which are most often lauded for one type—**internal validity**—and criticized for their lack of the other—**external validity**. Internal validity refers to the extent to which the effects of an experiment are actually due to the treatment, whereas external validity concerns the ability of a study to generalize beyond the subjects, settings, and treatments in that particular study—that is, to the idea that effects from the experiment also would be found in real life. Because external validity can be as big of a threat to experiments as internal validity is to observational studies,<sup>3</sup> this chapter begins with a discussion of external validity.

## ECOLOGICAL AND EXTERNAL VALIDITY

---

In broad strokes, external validity is the extent to which the results of an experiment can be generalized to other people, settings, and treatments. One of the concerns that affects external validity is how realistic (or artificial) the experiment is, called **ecological validity**. When an experiment reflects real-life circumstances or mimics the environment, providing subjects with an experience that is akin to what really happens, it is said to be ecologically valid.<sup>4</sup> Creating as much of a natural situation as possible enhances the chances of subjects responding as they normally would.<sup>5</sup> Quasi, natural, and field experiments are higher in ecological validity than are true or lab experiments. The challenge for lab experimenters is to maximize realism while still maintaining control. From chapter 2, we saw that Stanley Milgram and Philip Zimbardo worked hard to do just that. Without going so far as to construct a mock prison or shock-generating machine, today’s experimentalists still can achieve some degree of realism with a little extra effort and imagination. For example, running a political science experiment in an actual polling place, using real polling equipment and registered voters, helps enhance ecological validity.<sup>6</sup> Having subjects interact with a real teacher is more realistic than asking them to imagine an interaction from reading about it. Researchers in some disciplines find it easier to approximate real life than others. For example, studying whether getting a raise improves performance by bringing people into a lab, creating jobs, manipulating the reward, and then observing or measuring performance seems improbable for a business experiment. But a real-life setting affords no control, and people obviously cannot be randomly assigned to get raises. Having subjects read about a job, imagine themselves in that situation, and answer questions about what they might do may be the best available option. This is an example of the

trade-offs. Experimental economics has long had concerns over the artificiality of laboratory experiments, not only for setting and tasks that are unlike those in the real world but also for reactivity and demand effects, which were covered in chapter 2.<sup>7</sup> Researchers should understand the thinking in their particular field.

Creating realism often requires more work and creativity on the researcher's part but can be worth it in the results. I found this out firsthand when, as a graduate student, I helped conduct an experiment to determine if news stories that were framed differently would cause people to attribute blame for problems to society or individuals and to endorse different solutions. In the first attempt, my professor and I used real news stories but presented them to the subjects typed up on 8.5 x 11-inch paper. Beyond having a headline and byline, they did not look much like stories in a real newspaper. We found no effects. After some debate over what could have gone wrong, we decided to present the stories more realistically; instead of giving the stories to subjects on pieces of typing paper, we generated mock broadsheet newspapers with the manipulated stories embedded alongside other real news stories, photos, a nameplate, and page numbers. We printed them out on an oversized copy machine on paper that was 22-inches long and 13-inches wide, the same size as broadsheet newspapers. The same stories that produced no differences previously now showed the effects that theory had predicted.<sup>8</sup> All we did was change the appearance to be more realistic. It cost more and took longer, but the realism of the mock newspaper pages seemed to have been key. We did not make everything realistic; for example, our subjects were still reading the mock newspapers in a university classroom rather than at their breakfast tables, and it was a fictional newspaper name rather than a real one. But that one change seemed to be enough. A cheaper and easier compromise to the 13 x 22-inch broadsheet, I discovered, is what newspaper journalists call a "clip"—that is, a story that looks like it is cut out of a newspaper. I also find effects with clips, and they are easier to make. Today, researchers are more likely to have to create web pages or tweets. (Chapter 9 will discuss creating realistic stimuli.)

## GENERALIZABILITY

---

Realism is one of the keys to **generalizability**, or how well the effects of highly artificial experiments translate to real life.<sup>9</sup> In addition to how real it seems, external validity is also about whether results from a specific experiment will translate to other people, in other settings, at different times, and for different treatments.<sup>10</sup> When results generalize beyond the specifics of the study, they are said to be "**robust**."<sup>11</sup> Most experiments do not randomly sample from a population<sup>12</sup> the way surveys and other quantitative

"This is a public-health issue. Yes, we have newer medications and treatment methods, but that's not stopping the rise in cases," said Knight.

methods do. Random sampling, which allows every person in a population an equal chance of being chosen, results in a sample that affords the researcher the ability to say that the findings apply to the larger population with a certain level of probability. Most experiments do not randomly sample subjects, but there are other ways they achieve generalizability, including using people as close as possible as those to be generalized to. For example, experiments using twenty-year-old college students may not generalize to people on probation and parole. Findings about volunteers may not translate to people who do not volunteer.<sup>13</sup> What causes someone to donate to a health-related charity may be different for patrons of the arts.<sup>14</sup> Results from an experiment using accountants may not generalize to factory workers, and so on. The possibility that a treatment only works on the unique population used in the experiment is one concern of external validity.<sup>15</sup> One of the most serious limits to experiments is when they cannot be generalized beyond the specific people studied.<sup>16</sup> Campbell and Stanley give an example of a hypothetical researcher who tries to recruit ten different schools to participate in an experiment and is turned down by nine, saying, “This tenth almost certainly differs from the other nine, and from the universe of schools to which we would like to generalize, in many specific ways. It is, thus, nonrepresentative.”<sup>17</sup> Something about that school that led them to agree to be in the experiment might also cause the treatment to be effective, such as better teachers or a better location. For example, if the first four schools belong to one group of people, they cannot be generalized to a larger group that was not included in the study.<sup>18</sup> This leads to the obvious conclusion that researchers should use samples that are as similar as possible to the people or other units such as businesses, schools, or organizations the researcher intends to study—for example, using registered voters rather than those not registered, citizens from throughout the country rather than just one state, or accountants instead of students, if those first groups are the target populations to be generalized. This argues against what has become the rather routine use of students in subject pools; if the purpose of an experiment is to study decision making by military leaders or the type of public service announcements that best encourage parents to have their children vaccinated, then the use of sophomore advertising majors in the departmental subject pool is not a good choice. When subjects do not represent the population to which the results are supposed to apply, then external validity is low and the credibility of the results comes into question. If the population to which the research is aimed is advertising students, or even advertising professionals, for example, then the subject pool is more defensible, considering that advertising students will likely soon be advertising practitioners.<sup>19</sup> It may be easier to use students from the subject pool, but if those students do not approximate the population the experiment is designed to generalize results to, then it can be advantageous to expend more effort to find more representative subjects. Chapter 8 goes into more detail about the mechanics of selecting samples, and students in particular.

Campbell and Stanley describe twelve factors that jeopardize validity.<sup>20</sup> This chapter will not go into detail about all of them but will instead focus on a few of the more popular ways to combat low external validity. Occasionally, experimentalists will find that their experiments are being evaluated by journal reviewers who are experts in the topic but unfamiliar with experimental methods. This sometimes results in feedback that says the study should not be considered valid if the subjects were not randomly sampled, and should not be published. In those cases, it is helpful to be prepared with a knowledgeable, respectful response. This section is designed to help create experiments that avoid such criticism in the first place, and also to respond should it be given.

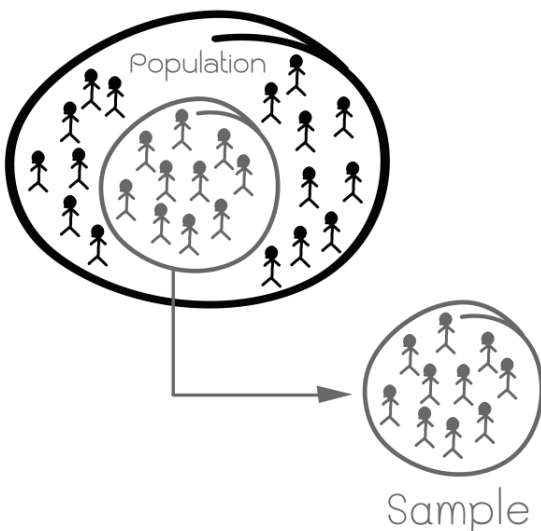
Recently, research has tackled the problem of generalizability in experiments and come up with ways to address the problem or at least estimate how likely the results are to accurately represent a particular population beyond the study subjects.<sup>21</sup> Some include different methods for selecting subjects, while others create measures from outcome variables or covariates to estimate how representative of the population the experimental sample is.<sup>22</sup> Some of these techniques can be used when selecting samples, and others are used after data are collected.<sup>23</sup> These are fairly advanced techniques, which this textbook will not discuss in detail; rather, this chapter covers the basic issues and solutions.

## For Review-No Commercial Use(2023) Random Sampling

Just as random assignment is the gold standard for achieving equivalent groups in experiments, the gold standard for achieving generalizability is random sampling.<sup>24</sup> As we have known since 1936 when *Literary Digest* magazine predicted the wrong winner of the U.S.

presidential election while George Gallup's polling organization predicted it correctly,<sup>25</sup> random sampling is the most efficient and effective way to estimate some characteristic of an entire population from a smaller sample of it. But you rarely see it used in experiments.<sup>26</sup>

First, we should make clear the difference between random *assignment* and random *sampling*. Random sampling is the process of selecting subjects for a study so that they adequately represent a larger population. It ensures that each person, site, or other unit from a population has an equal chance of being chosen, and is cheaper and easier than measuring a whole population,



especially when it would be impossible to do so. Random assignment, by contrast, is the act of deciding which subjects get a particular experimental treatment or are put into a particular group. Random assignment is typically done with **convenience samples**, or those subjects who are easily available, rather than with subjects who are randomly sampled. The purposes of each are quite different. “Random assignment . . . facilitates causal inference by making samples randomly similar to *each other*, whereas random sampling makes a sample similar to *a population*.”<sup>27</sup> In the case of experiments, it is not as important that the subjects look like the people who are not in the study, but that the subjects in the different groups of the study look like each other.

### Two-Step Randomization

If each of these kinds of randomization is the gold standard, then the use of both in combination surely should be platinum. In fact, some experiments do use a **two-step randomization model**, where subjects are first randomly sampled from a population and then randomly assigned to a treatment group,<sup>28</sup> but that is rare. One of the most obvious reasons is that researchers would need a complete and up-to-date list of all the members of a population in order to randomly sample from. The larger the population, the less likely this is to exist. For example, no complete list exists of all parents of school-age children in the United States. In fact, we have no idea if private schools are being homeschooled. And while there are lists of registered voters, there are not lists of *unregistered* ones. Constraining the population to be generalized makes random sampling more feasible in some cases; for example, it would be possible to obtain a list of all school students in a certain school district, but the narrowing of the sample raises the same questions as before—do effects from one school district generalize to others? If a researcher could draw a random sample from across the United States, it would be difficult, if not impossible, to administer some of the treatments experimentalists use to all these people scattered all across the country. Researchers would have to travel around the country or bring the subjects to the lab, both of which are expensive, or rely on others to give the manipulation, which jeopardizes internal validity because subjects respond differently to different experimenters who may even deliver the stimuli in a different way. Besides these logistic considerations, there are budget realities. For example, a database of all working U.S. journalists *is* available commercially, but it costs thousands of dollars. That is understandable when considering how much time and effort it takes to compile such information. Most researchers do not have this kind of time or money. Researchers should weigh the benefits against the costs.

### Nonresponse Bias

In situations where a population list is easier to acquire, experimentalists then run into the same kinds of problems that survey researchers do when recruiting people. To estimate

a population requires a lot more participants than it does to find an effect; for example, 400 might be needed for generalizing instead of the forty that is more common to experiments that are only looking for effects.<sup>29</sup> Therefore, if an experimenter also wants to generalize, more subjects are needed, which costs more. Also, not everyone who is randomly chosen will agree to be in a study; poll and survey researchers deal with this in one of their biggest external validity concerns—**nonresponse bias**, which happens when people do not answer a question or refuse to participate. People who do not respond can be very different in unknown ways from those who do, such as the schools that refused to participate in Campbell and Stanley’s fictional example. Nonresponse bias carries threats to generalizability, and response rates have been dipping into the teens and below for years. There also is the worry about **attrition**, or subjects who drop out before completing a study. It does not make sense for an experimental researcher to go to all the time, expense, and trouble to obtain a list of a population to randomly sample from, only to discover that not enough people will agree to participate to ensure generalizability anyway. When it *is* possible to obtain a list of all units in the study, such as Malesky and colleagues<sup>30</sup> did with firms in Vietnam for an online experiment of bribery, and then randomly sample them before randomly assigning them to conditions, response rates must be calculated and reported, and a discussion of response bias presented (see Study Spotlight 5.1, Generalizability and Representativeness). This method of random sampling did not make the generalizability question disappear. The authors provided a table comparing the characteristics of their sample to those of the population

These issues make random sampling “an inadequate solution” to the problem of generalization from experiments.<sup>32</sup> In addition, some businesses, government agencies, or other organizations will not allow their employees or members to be randomly sampled; they may feel it is unfair to give some an opportunity that others do not have, among other reasons. If researchers are offering a career development seminar as the manipulation, for example, employers may be reluctant to tell some employees they cannot participate and invite charges of favoritism and discrimination. Because random sampling requires “a high level of resources and a degree of logistical control that is rarely feasible,”<sup>33</sup> most experimentalists use other sampling methods that are accepted throughout all fields of scientific experimentation. These sampling methods are the topic of chapter 8. Instead, to give us more confidence in the generalizability of experiments, researchers use other approaches, described next.

## Representativeness

The goal of random sampling from a larger population is to ensure that the smaller sample looks like the larger population—that is, that it is representative. In the absence of



## STUDY SPOTLIGHT 5.1

### Generalizability and Representativeness

Next is an example of a discussion of generalizability of the experiment. The authors included a table comparing the sample to the population for representativeness.

**Malesky, Edmund J., Dimitar D. Gueorguiev, and Nathan M. Jensen. 2015. "Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment." *American Journal of Political Science* 59 (2): 419–439.**

Appendix 4: Characteristics of Provincial Competitiveness Index Sample and Census Data in 2010

Foreign invested (3,888)			Domestic enterprises (19,363)		
<u>Legal form of investment</u>	<u>Weighted PCI</u>	<u>GSO</u>	<u>Legal form of investment</u>	<u>Weighted PCI</u>	<u>Tax</u>
100% Foreign-directed enterprise	84.35%	82.95%	Sole proprietorship	16.2%	19.4%
Joint venture with a Vietnamese private	4.84%	16.36%	Limited liability	54.5%	59.1%
Joint venture with a Vietnamese SOE	4.55%		Joint stock	27.6%	21.4%
Registered as a domestic company	2.52%	0.46%	Joint stock with shared listed on stock exchange	1.1%	NA
Domestic company w/overseas VN capital	0.61%		Partnership and other	0.7%	0.0%
Other	3.13%	0.23%			
<u>Sector</u>	<u>Weighted PCI</u>	<u>GSO</u>	<u>Sector</u>	<u>Weighted PCI</u>	<u>Tax</u>
Industry/manufacturing	64.59%	59.44%	Industry/manufacturing	30.2%	34.5%
Construction/infrastructure investment	4.09%	4.72%	Construction/infrastructure investment*		
Service/commerce/finance	29.33%	28.94%	Service/commerce/finance	64.6%	62.2%
Agriculture/forestry/fishing	2.26%	5.97%	Agriculture/forestry/fishing	4.0%	1.9%
Mining/natural resource exploitation	0.94%	2.13%	Mining/natural resource exploitation	1.2%	1.4%
<u>Size of labor force</u>	<u>Weighted PCI</u>	<u>GSO</u>	<u>Size of labor force</u>	<u>Weighted PCI</u>	<u>GSO</u>
Less than 5	2.92%	4.18%	Under 5	12.1%	23.36%
5 to 9	5.99%	6.79%	5 to 9	24.1%	35.63%
10 to 49	31.79%	29.67%	10 to 49	41.9%	33.22%
50 to 299	31.35%	30.95%	50 to 200	14.9%	6.11%
300 to 399	6.38%	7.64%	Over 200	7.1%	1.7%
400 to 499	7.26%	7.09%			
500 to 999	7.17%	6.88%			
1000 and over	7.13%	7.81%			
<u>Licensed investment size</u>	<u>Weighted PCI</u>	<u>GSO</u>	<u>Licensed investment size (Total assets, BVND)</u>	<u>Weighted PCI</u>	<u>GSO</u>
Under 0.5 BVND (\$25,000 USD)	2.52%	2.25%	Under 0.5 BVND (\$25,000 USD)	10.9%	8.9%
From 0.5 to under 1 BVND (\$50,000 USD)	1.39%	2.17%	From 0.5 to under 1 BVND (\$50,000 USD)	17.0%	13.5%
From 1 to under 5 BVND (\$250,000 USD)	15.85%	12.75%	From 1 to under 5 BVND (\$250,000 USD)	42.8%	49.6%
From 5 to under 10 BVND (\$500,000 USD)	8.75%	11.71%	From 5 to under 10 BVND (\$500,000 USD)	12.7%	13.4%
From 10 to under 50 BVND (\$2.5 Million USD)	35.14%	36.04%	From 10 to under 50 BVND (\$2.5 Million USD)	11.9%	11.5%
From 50 to under 200 BVND (\$10 Million USD)	23.13%	22.83%	From 50 to under 200 BVND (\$10 Million USD)	4.8%	3.2%
From 200 to under 500 BVND (\$25 Million USD)	7.62%	7.29%	From 200 to under 500 BVND (\$25 Million USD)		
Above 500 BVND (\$25 Million USD)	5.61%	4.97%	Above 500 BVND (\$25 Million USD)		
<u>Major customer</u>	<u>Weighted PCI</u>	<u>GSO</u>	<u>Major customer</u>	<u>Weighted PCI</u>	<u>GSO</u>
Export directly or indirectly	55.00%	66.8%	Export directly or indirectly	11.7%	NA
Foreign individuals or companies in Vietnam	24.51%	16.2%	Foreign individuals or companies in Vietnam	9.9%	NA
Sold domestically to SOE	3.52%	2.8%	Sold domestically to SOE	14.8%	NA
Sold domestically to state agency	1.42%	0.9%	Sold domestically to state agency	20.3%	NA
Sold domestically to private individuals	15.55%	13.0%	Sold domestically to private individuals	43.4%	NA

Note: This table compares data on the nationally weighted sample of domestic and foreign firms from the PCI to the data collected from the National Tax Authority (Tax) and General Statistical Office (GSO) Enterprise Census. Weighted PCI is the PCI survey sample, but weighted by provincial share of enterprises to create a nationally representative sample. General Statistical Office (GSO) data available at ([www.gso.gov.vn](http://www.gso.gov.vn)) and GSO Enterprise Census (2009) available at ([http://www.gso.gov.vn/default\\_en.aspx?tabid=515&idmid=5&itemID=9775](http://www.gso.gov.vn/default_en.aspx?tabid=515&idmid=5&itemID=9775)). NA = Not Available for 2010. \*Tax Authority data does not disaggregate construction firm from manufacturing. The PCI data records 15 percent construction.

PCI = Provincial Competitiveness Index

BVND = Billion Vietnamese Dollars

SOE = state-owned enterprise

VN = Vietnamese

Source: Survey data from Vietnam PCI 2010 Report ([www.pcivietnam.org](http://www.pcivietnam.org)); and GSO Enterprise Census 2009 ([www.gso.gov.vn](http://www.gso.gov.vn))

(Continued)

(Continued)

“Our final sample is composed of 19,363 domestic firms and 3,888 FIEs, which registered after 2000 and are located throughout the country’s 63 provinces. In all three years, the sample frame for selection was the list of registered domestic firms and FIEs in the General Tax Authority database of registered operations. The survey response rate was about 30% for domestic operations and 25% for FIEs, much higher than rates commonly reported in the international business literature (White and Luo 2006), but still small enough to create concerns about reliability (Dillman et al. 2002). As a result, it is reasonable to ask whether nonresponse creates selection bias that might affect our conclusions (Jensen, Li, and Rahman 2010). In Appendix 4 in the supporting information, we compare the PCI data to available information from the General Statistical Office’s Enterprise Census and Tax Authority Databases in 2012, showing that PCI data in a given year reflect observable characteristics of the national population and therefore offer a highly accurate depiction of foreign and domestic investors in Vietnam.”<sup>31</sup>

the ability to randomly sample, taking precautions to ensure that the sample adequately represents the rest of the population is one step that gives us more confidence in the external validity of the results of experiments.<sup>35</sup> Shadish, Cook, and Campbell discuss the various ways that the *non-representative* has been used in experiments, including as a miniature of the population, saying this is used to “describe samples that are representative in a purposive sense.”<sup>36</sup> The goal is to know the key characteristics of the population and ensure that the sample adequately represents as many of these key characteristics as possible. Many experiments take pains to compare their sample’s characteristics to those of the target population—for example, a national census.<sup>37</sup> It is also important to point out where samples differ from the population.<sup>38</sup> It is acceptable for the sample not to be in the exact proportions of the population on some demographics, as similarity to some variables but not all of the key characteristics are all that is required.<sup>39</sup> For example, Klaiman and colleagues’ sample mirrored the census in gender, age, income, and rural and urban regions, but was slightly more educated than the population.<sup>40</sup> Researchers should be especially concerned about variables that affect the outcome variable.

Regardless of how good the fit, this type of generalizing is never backed up by statistical logic yet remains widely used in experiments.<sup>41</sup> Verschoor et al.’s<sup>42</sup> study of economic risk taking specifically tests the question of how well experiments reflect real-world behavior. In this study of farmers in Uganda, the researchers use a representative sample and bill their study as “a test of the external validity of the experimental method,”<sup>43</sup> finding that indeed, this experiment that used a representative population was externally valid if enough care was taken with sampling methods and other validity concerns, such as minimizing confounding variables.

## MORE ABOUT . . . BOX 5.2

### Experiments With Random Sampling and Random Assignment

True experiments that use both random sampling and random assignment are rare. While it is not recommended as the only way to generalize to a population, Shadish, Cook, and Campbell do say, “We unambiguously advocate it when it is feasible.”<sup>34</sup>

“Feasible” usually means a population that is constrained to a manageable level, such as residents of one city or state, or when a list of an entire population exists, such as for members of a licensed profession. Random sampling is also facilitated when databases are publicly available, researchers enlist the aid of officials with access to information about a population, or can purchase databases; for example, the company Cision tracks all print, broadcast, and online journalists in the United States, and also top social media influencers (<http://www.cision.com/us/pr-software/media-database/>).

Two examples of experiments that use two-step randomization are summarized next:

**Lu, Fangwen, Jinan Zhang, and Jeffrey M. Perloff. 2016. “General and Specific Information in Deterring Traffic Violations: Evidence From a Randomized Experiment.” *Journal of Economic Behavior & Organization* 123: 97–107.**

This experiment conducted in Tsingtao, China, used the city’s database of registered vehicles, which also listed cell phone numbers of the owners. The researchers conducted a census of 25,443 car owners. The researchers enlisted the cooperation of the Tsingtao Police Department to provide the data. The study also randomly assigned the registered vehicle owners to one of several treatment conditions where they were sent text messages with varying content about traffic violations they had received, a general safety message, or a control group that received no message. They then tracked to see which safety message resulted in drivers incurring fewer violations in the future.

**Shi, Ying. 2016. “Cross-Cutting Messages and Voter Turnout: Evidence From a Same-Sex Marriage Amendment.” *Political Communication* 33 (3): 433–459.**

This study used publicly available databases of voter registration and voting history from the state of North Carolina as the population, and then drew a random sample of registered Democrats and Republicans who then received direct-mail postcards with messages either for or against their party’s position on same-sex marriage to see if cross-cutting views stimulate or discourage political participation, measured by whether they voted in the election after receiving the postcards.

There are various ways to increase the chances of a sample being representative—for example, oversampling of certain hard-to-obtain populations and use of multiple sites, which should add diversity.<sup>44</sup> The use of online survey methods to conduct experiments, frequently called **survey experiments**, also has increased the ability of experiments to use representative samples,<sup>45</sup> as has the use of panels of survey respondents recruited by sampling firms.<sup>46</sup> These will be discussed in more detail in chapter 8.

When experimental samples do not accurately represent populations, the paper's discussion section should include an acknowledgment of this limitation and an explanation of why this is not problematic or should not invalidate the findings.<sup>47</sup> (See How To Do It box 5.3 for examples of how some experimentalists have handled this.) A good practice is to always say something like “findings should be interpreted with caution” if the sample is not representative of the population, and to be careful to talk about findings from “these subjects” or “the people in this sample,” “our subjects,” or similar language that does not imply extrapolating beyond those in the sample.

## CAUSE AND EFFECT

It can also be a good idea to point out in the article the fact that experiments are not designed to be generalizable but are intended to find effects if they exist. In most experiments,

### HOW TO DO IT 5.3

#### What to Say When It Is Not Representative

#### For Review-No Commercial Use(2023)

When samples are not representative of the target population, experimentalists are up front about it but defend the results on various other grounds. Here are some excerpts of such discussions from a few studies:

**Kim, Mirae, and Gregg G. Van Ryzin. 2014. “Impact of Government Funding on Donations to Arts Organizations: A Survey Experiment.” *Nonprofit and Voluntary Sector Quarterly* 43 (5): 910–925.**

“Compared to the U.S. population, there are relatively high proportions of females and Whites in the panel, a feature that Berrens, Bohara, Jenkins-Smith, Sliva, and Weimer (2003) find to be common in other voluntary online panels in the United States. Although this limits the external validity of studies using online panels, there is some evidence to suggest that, at least for some topics, this type of voluntary panel can approximate survey results based on probability sampling of the U.S. population. Although not representative of the general population, the participants in our study clearly appear interested and supportive of the arts and thus perhaps more closely resemble the potential donor community for many U.S. arts nonprofits” (pp. 916–917).

**Dunlop, Claire A., and Claudio M. Radaelli. 2016. “Teaching Regulatory Humility: Experimenting with Student Practitioners.” *Politics* 36 (1): 79–94.**

“Neither of the groups are representative samples of public sector administrators or MPA students. This is not problematic, however, given that our purpose is to illustrate the utility of experiments in the classroom in helping student-practitioners critically reflect rather than make any empirical contribution to the illusion of control literature. The utility of data gathered is pedagogic rather than research-oriented” (p. 82).

representing a larger population is not the goal.<sup>48</sup> Rather, experiments are good for studying theoretically important relationships. This is a fundamental misunderstanding about experimental research that the public, media, and researchers using other methods occasionally make.<sup>49</sup> Researchers can get caught up in issues such as generalizability and lose sight of the fact that the purpose of an experiment is to establish causality—that is, to discover if some treatment variable is the cause of some outcome or effect. The purpose of laboratory experiments is not to generalize to a particular population.<sup>50</sup> And in this regard, the issue of random assignment is far more crucial; more about that in chapter 7.

Surveys, content analyses, and other methods have the opposite problem that experiments have; if sampled correctly, they can generalize to the population, but they cannot infer the cause of the relationships they find.<sup>51</sup> These kinds of studies should be careful to talk about “associations” or “relationships” rather than declaring or implying “cause.” Generalizability is more important for survey researchers, who are interested in obtaining the attitudes and opinions of people. It is more reasonable to ask if those attitudes and opinions are widely held by others. But for experiments, aimed at determining cause and effect, that question takes a backseat to whether the outcome was confounded by something other than the manipulation or treatment.

## For Review-No Commercial Use(2023) LOGICAL INFERENCE

---

There can also be a misunderstanding about the kind of inferring that experiments attempt. Rather than making a **statistical inference**, which means that a confidence level or interval can be mathematically calculated to determine the fit between the sample and larger population, experimentalists rely on **logical inference**, the idea that reasonable conclusions can be drawn on the basis of common sense. For example, if an experiment tests some function that is relatively unaffected by race, gender, or where one lives, among other things<sup>52</sup>—for example, the cognitive process of opinion formation—then it is appropriate to draw inferences to others not in the sample. Basil gives an example of hardwired physiological responses to photographs taken from a close range versus photographs that show something or someone farther away, saying arousal to such images should be the same regardless of who is in the sample.<sup>53</sup> It is reasonable to assume that such processes should be similar for others not in the sample if there is no evidence to the contrary. For example, if college students were more afraid of scary close-up shots than scary far-away shots, why would we not expect the same response among noncollege students? Similarly, media effects such as agenda setting, priming, and framing have been shown to work the same way with a variety of populations all around the country and even the world.<sup>54</sup> If, however, what is being studied changes with various conditions—for example, in cultures that have

state-run media—we would not generalize to them from conclusions made in different conditions, such as in free-press societies. Unless a phenomenon is different for different people, the findings should hold for everyone. The real concern for experiments is whether the sample is appropriate to what is being studied, not as much as how it was drawn.<sup>55</sup>

Finally, to dismiss experiments based on the generalizability of the samples used would be to reject findings in medicine, psychology, chemistry, and physics, among other disciplines, that have made valuable contributions to science and everyday life based largely on nonrepresentative samples.

## REPLICATION

---

Rather than concentrating so much on how samples are drawn or how well they mimic a larger population, experiments rely primarily on **replication** to make generalizations about their conclusions.<sup>56</sup> Replication is the ability to repeat findings from an experiment with different subjects, under different conditions, at different times and places, with different issues, variables, etc. If consistent effects are found in different studies, then there is more confidence that those findings represent something real and generalizable rather than being caused by chance, bias, or error.<sup>57</sup> Replications are “purposeful duplications of the methods used that specifically assess the veracity of previous research results.”<sup>58</sup> Some professions that rely on evidence-based practices are especially concerned with replication to make sure results are reproducible before instituting costly changes in practice.<sup>59</sup> Political science began a replication movement in the 1990s,<sup>60</sup> and social psychology has recently renewed its own, begun in the 1970s.<sup>61</sup> Unfortunately, replication is not as prevalent as calls for more of it; for example, in special education, only 0.5% of articles sought to replicate previous findings,<sup>62</sup> up from 0.13% in education in general.<sup>63</sup> Psychology journals are among the highest, at 1.07%,<sup>64</sup> which still seems pretty low.

Social scientists working in every method have often heard that a single study never proves anything; this is equally true of experiments. Researchers also advise caution with language—the “p” word (proves) should not be used even with results that are replicated; rather, good phrases include things like “this increases our confidence in the ability to generalize,” or “with replication, we are in a stronger position to make generalizations or causal inferences.”

### Self-Replication

Researchers can strengthen the generalizability of their own findings by replicating their own work, called **self-replication**. In some fields, this is required to demonstrate the results were not due to some quirk. Replicating one’s own study with a

second study in the same article is one way replication is achieved. The number of experiments in one article was up to three by 1998 for one social psychology journal.<sup>65</sup> This also fits into a systematic program of work that is encouraged by academic departments. For example, a program of work on moral judgment in journalism started with an experiment that manipulated whether journalists saw photographs or not when making ethical decisions and found that the photographs improved the quality of their moral judgment.<sup>66</sup> The main results were replicated in a second experiment in the same paper. Those experiments raised more questions about the process of journalists' moral decision making, so research continued to explore that topic, following it with three studies of whether race made a difference,<sup>67</sup> and another one that used audiences rather than journalists.<sup>68</sup> All these experiments tested whether photographs improved moral judgment and found that they did. Consistent results across five studies using students, professionals, and audiences of different races, in different parts of the country, at different times, gives us confidence to make a more firm assertion about the ability of photographs to elevate the quality of reasons used to make ethical decisions.

### Exact vs. Conceptual Replication

#### For Review-No Commercial Use(2023)

One important thing to note is that none of these follow-up studies were **exact** or **direct replications** of the first one—that is, they did not duplicate the original methods and procedures exactly. That type of replication is not as valued or recommended in the social sciences as it is in other fields such as medicine.<sup>69</sup> Exact replications can even be impossible in social science, as *something* must be different, such as the subjects.<sup>70</sup> Rather, replications are expected to also *extend* a study in some theoretically meaningful way. These are called **conceptual replications**,<sup>71</sup> or reproductions,<sup>72</sup> with key features deliberately varied, such as the setting, subjects, variables, or interventions. Conceptual replications test the theory of a study, while direct or exact replications test its operationalization. Conceptual replications extend an experiment's generalizability,<sup>73</sup> thus, the kind this chapter focuses on. In the previous example, examining how race communicated via photographs affected moral judgment of journalism students was a first step in extending the findings about effects of photographs. The study after that did the same but by investigating in-group/out-group processes using Black journalism students, because race was expected to work differently for minorities, which it did. The next one extended it from journalism students to professional journalists of three different races, and also examined the mediating role of empathy. The last one moved away from journalists to news audiences and examined how cognitive elaboration worked. These studies represent a systematic progression of replication by logically expanding beyond one type of subject and also incorporating different concepts theorized to be



causal mechanisms in the process of moral judgment when photographs are involved. Causal mechanisms are the variables or pathways through which some outcome occurs; they are what links some cause to an effect. In this case, race, empathy, and elaboration were the causal mechanisms that led to better moral judgment (the outcome or effect) from seeing photographs (the cause). Recall in chapter 1 the discussion of explanations for why some effect occurs. A good example of replication by systematic extension of findings to different populations, in different settings at different times, and the incorporation of different theoretical concepts as causal mechanisms is found in Stanley Milgram's book about his series of experiments on obedience to authority, which was covered in chapter 2.<sup>74</sup>

## Multiple Experiments

Many social science experiments are reported as single studies, but some disciplines are more apt to report multiple experiments in one article. For example, social psychology is moving toward multiple studies in one paper as a means of evidence of replicability.<sup>75</sup> Multiple experiments are also becoming more common in communication, with successive experiments systematically eliminating alternative explanations.<sup>76</sup> Fiske recommends researchers publish multiple studies in one paper.<sup>77</sup> And while it may seem more helpful to one's career trajectory to publish more articles by splitting up studies into one per paper, it aids the cause of generalizability to report multiple experiments in one paper, as consistent findings provide stronger reinforcement for conclusions. One example is an article by Schweizer,<sup>78</sup> who reports in one paper the results of two experiments designed to assess the effectiveness of teaching students a skill that promotes idea generation and discussion. The second experiment replicated the first one's results and extended them by determining how many classroom sessions were necessary for students to learn the technique.

It also makes sense to report more than one experiment in a single paper when one study raises questions or plausible alternative explanations for the results and a second experiment is needed to answer those questions or rule out the alternative explanations. By itself, the paper may not pass muster with journal reviewers without the additional experiment to explain findings and answer questions. An example of multiple experiments in one article is a paper that reports two experiments on the effects of harm and disgust on moral judgment in politics.<sup>79</sup> The second experiment was devised to test alternative explanations that arose in the first, in addition to replicating those results. It also introduced new political issues and so replicated results for three different issues. For more examples of replications, see More About box 5.4.

## External Replication

Replicating one's own work does help with making it generalizable, but it is also desirable to have results replicated by independent researchers—that is, by someone other than yourself, your coauthors, or even others at your own institution. This is known as **external replication**. Confidence is increased if an experiment can be replicated by outside researchers, independent of the first set, as it helps confirm that the previous results were not the result of a mistake, poor-quality work, or even outright fraud.<sup>80</sup> Some studies have found that when one or more of the authors on a replication were also involved in the original study, they were significantly more likely to reproduce the original findings than when outside researchers attempted replications.<sup>81</sup> Of course, there are plenty of reasons for this that do not point to nefarious activity. Nor does one failed replication necessarily mean the effect was false or the study a fraud. Rather, results are considered holistically, and when multiple replications show similar patterns or are close to the original study's results, they are considered successful.<sup>82</sup> When an external replication fails to reproduce the original's results, it could be because of too small a sample size, bad measures, experimenter effects, execution, or many other reasons. Rather than spelling doom, it encourages more study. This may be a sign that something new or interesting has been found; for example, if a political strategy has consistently worked but is suddenly not, this could be a sign that strategy should be updated rather than a sign that the experiment was flawed. Nonstatistically significant finds are still meaningful and should be reported.<sup>i</sup> It is too long to report here, but to see how researchers responded when external researchers failed to replicate their original findings, read Ebersole et al. (2017).<sup>83</sup>

In fact, it is common for results not to be replicated or to reproduce some results but not others. Some studies have found replication rates of 50% or lower.<sup>84</sup> In the example of research on photographs' ability to improve moral judgment, after five successful conceptual replications, one did not show the same thing.<sup>85</sup> That paper proposed theoretical reasons, backed up by empirical evidence, for why this happened. One thing that is *not* ethical to do is to fail to report or not attempt to publish failed replications. Such a lack of transparency or attempt to cover up unwanted results can result in the end of a promising research career, as some have taken it upon themselves to police the profession.<sup>86</sup>

Scientific knowledge is cumulative; it accrues over time across related studies. Thus, conceptual replication is important not just to generalizability but also to advance our understanding of important human phenomena. Social science research is about discovering as much as possible about human processes, not about supporting one's hypotheses at all costs.

<sup>i</sup>I would like to thank an anonymous reviewer for suggesting the addition of these last two observations.

## MORE ABOUT . . . BOX 5.4

### Examples of Replication Studies

**Krupnikov, Yanna, and Spencer Piston. 2015. "Accentuating the Negative: Candidate Race and Campaign Strategy." *Political Communication* 32 (1): 152–173.**

This study experimentally manipulated race and tone to see how voters responded to Black and White candidates' negative campaign ads. It reproduced the results of previous work and extended it by studying the interaction of race and negativity; the two had been considered independently in previous studies. The authors did a second experiment to test alternative explanations that arose in the first study and replicated their own results as well. This experiment used a nationally representative sample embedded in a survey.

**Camerer, Colin F., Anna Dreber, Eskil Forsell, Ho Teck-Hua, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics. *Science* 351 (6280): 1433–1436.**

The reason why this article has so many authors will be apparent once you learn that this study replicated the most important finding from eighteen different laboratory experiments in economics. They were able to reproduce findings on eleven of them (61%). They note the limitations include a smaller number of studies to be replicated than many other reproducibility projects. Details of all eighteen studies are not in the article but on a companion website.

**Nazarian, Deborah, and Joshua M. Smyth. 2010. "Context Moderates the Effects of an Expressive Writing Intervention: A Randomized Two-Study Replication and Extension." *Journal of Social and Clinical Psychology* 29 (8): 903–929.**

The two experiments reported in this article tested the effects of expressive writing by changing where subjects wrote—at home or in a lab—and the authority of the experimenter via how formally he or she was dressed. This study also speaks to the issue of ecological validity, or generalizability outside the laboratory, as it found that students did better when writing in a lab and adults did better when writing at home. The use of students in one experiment and adults in the other also aided the cause of generalizability to different populations.

One of the main goals of a well-written methods section is so others may replicate the study. Methods sections are like the recipe for an experiment, with enough detail that others should be able to reproduce a study without having to contact the original author. When methods and protocols are clear, complete, and precise, it is more likely that other researchers will be able to successfully replicate results because they will have all the necessary elements exactly

the same. Another movement currently in progress is to make freely available one's data and the coding used to analyze it so that others may easily check the work. It has always been an ethical tenet that researchers make their data available to others who request it. The web is making this practice more widespread, as researchers can simply download the data rather than make requests of other researchers. More and more journals are encouraging authors to include not only the datasets and code but also stimuli and other components on their supplemental materials website. Some researchers even include them on their own websites or on academic social media sites like Researchgate.net and Academia.edu. While a few journals require it, most are still voluntary. Similarly, a few journals are including regular sections devoted to replications.<sup>87</sup>

## Triangulation

Another means of establishing the validity of some finding is by using both qualitative and quantitative methods. This is also important to generalizability. If findings from a lab experiment are also the conclusions of interviews, ethnographies, and other methods, that can have important implications for real-world relevance.<sup>88</sup> Fiske and others recommend field experiments as replications of lab experiments because of their high external validity.<sup>89</sup>

A final word on external validity concerns the use of pretests, a discussion started in chapter 4 that will be revisited later. In addition to the pros and cons outlined previously, experimental designs without pretests are preferred for reasons of external validity.<sup>90</sup> Because of the interaction of testing and treatment, exposing subjects to the test twice may sensitize them to the topic or have the opposite effect and diminish their sensitivity to it, reducing a treatment's effectiveness. Thus, results with subjects who have been tested more than once may not generalize to subjects who have not had multiple tests.<sup>91</sup>

This is not a complete list of all threats to external validity. Others expand on these basic threats with, for example, lack of message variance and message repetition, which involves giving subjects more than one version of a stimulus for each treatment level. By using just one stimulus, the only conclusion that can be reached is that exact stimulus had an effect, not that all stimuli of that type have an effect.<sup>92</sup> This will be discussed in more detail in chapter 9.

## INTERNAL VALIDITY

All the concerns about generalizability and ecological validity discussed thus far are of no consequence if the findings from an experiment are not accurate because of a problem with internal validity. Campbell and Stanley call it the “basic minimum” requirement

for any experiment.<sup>93</sup> Internal validity is the primary strength of true experiments; no other method offers researchers as high a degree of internal control—that is, control over the causal mechanisms between conditions.<sup>94</sup> Internal validity goes to the heart of the cause and effect relationship—that what the experiment found to be the cause is actually responsible for the outcome rather than something else. In the trade-off between external and internal validity, experimenters usually choose internal validity.

### Three Basic Criteria

The three basic criteria for an experiment covered in chapter 1 speak directly to internal validity, as reviewed briefly: (1) that the cause comes before the effect, (2) that the cause and effect are related, and (3) that there are no other plausible alternative explanations for the effect. The first two are barely ever discussed in laboratory experiments; which came first is more a concern of field and natural experiments. For example, whether unemployment causes inflation or inflation causes unemployment is a chicken-and-egg-like question. If you cannot definitively say which came first, it is not internally valid. This is mostly a concern with Campbell and Stanley's three pre-experimental designs, quasi or natural experiments. But in laboratory or true experiments, researchers should ensure that subjects get the treatment or manipulation before measuring the outcome. The second criteria for internal validity, that the cause and effect be related, is illustrated by the example in chapter 1 about whether storks bring babies. Except perhaps to those living in Denmark in the 1930s, a relationship between storks and babies is obviously not logical, nor is it supported by any theory.<sup>95</sup> This is also why experimenters do not test hunches but must first conduct a literature review and establish theoretical and empirical links between any proposed cause and effect. Performing that task, detailed in chapter 3 on theory and literature, should quickly disabuse a researcher of spurious notions about cause and effect. Rather, it is the third criterion that occupies much of experimentalists' time: ensuring there are no confounding factors or plausible alternative explanations that account for the effect. Internal validity in experiments is achieved when there is tight management of all possible confounds and alternative plausible explanations, minimal bias and error, and strong controls in place. This is the key task that experimentalists will encounter related to internal validity.

While there are an unlimited number of confounds and other plausible explanations for an outcome in experimental studies, there is, unfortunately, no comprehensive list of them. Reviewers are expert at pointing out what they are, though. The trick is to identify them ahead of time, which is accomplished by thoroughly reviewing the literature and also by applying one's own expertise and common sense. One example of a possible confound in the studies on photographs' effect on moral judgment<sup>96</sup> was whether the people in the photographs were close up or far away. A thorough review of the literature resulted

in evidence from the study of proxemics in nonverbal behavior, which showed that being close to others makes people believe they are more persuasive, credible, and sociable than does being farther away.<sup>97</sup> Had the studies not controlled for proximal distance by using pictures that were the same in terms of shot distance, the real cause of the bump in moral judgment could have been due to how close or far away people in the photographs were rather than what was being manipulated. Another example of this is in experiments that control for message length; we know that giving people more information can affect their knowledge levels, attitudes, and other outcome variables, so it is standard practice to keep the length of manipulated messages as similar as possible to control for this confound.<sup>98</sup>

## Random Assignment

The gold standard of random assignment also aids the cause of internal validity by making comparison groups as similar as possible. Thus, there should not be anything inherently different about one group that caused a treatment to work on one group and not the other. Therefore, we can assume that it was the treatment that caused the outcome. The benefits of random assignment cannot be stressed enough, and this will get its own chapter in chapter 7.

## Confounds For Review-No Commercial Use(2023)

When possible confounds are only raised after a study has been conducted, this provides fodder for replications that extend a study, as described earlier in the external validity section. Such conceptual replications mainly for purposes of external validity also aid the cause of internal validity by ruling out confounding variables as plausible explanations. When a possible confound is identified, it is sometimes possible to rule out the plausible alternative explanation by doing additional analyses with the data the researcher already has. For example, a reviewer once questioned the way religiosity was measured when it was made up of three questions. The reviewer disagreed that one of the questions was a true measure of the concept and suggested it might have confounded the results. To see if this concern was indeed serious, the authors reanalyzed the data, eliminating the offending question and using only a two-item measure of religiosity. The results were the same. This satisfied the reviewer and the article was published. Another time, a reviewer was concerned that subjects had tried only to support their preexisting ethical decisions rather than seriously grapple with other points of view to come to a considered judgment as was proposed. The concerns were addressed by going back through the open-ended responses subjects had given and demonstrating that, indeed, they had considered arguments opposite to their own by counting the self-supporting and counterfactual arguments and reporting this, with examples. For the explanation that ended up in print, see How To Do It box 5.5.

## HOW TO DO IT 5.5

### Explaining Experimental Validity and Confounds

In the review process for this study, reviewers suggested the results might have been confounded by variables such as different degrees of violence in the images, the emotions they created in the viewers, or the seriousness of the crime depicted. In the response, the authors invoked both the controlled nature of experiments and also previous empirical evidence. Here is what ended up in the article:

“Because this was a controlled experiment, we can be sure that this effect was due to the only thing that was altered—the format of the visual. We can say for certain that it was not emotion, violent images, seriousness of the crime or some other variable because those were the same for all the subjects—all subjects saw all three stories. The videos were the same, with the only difference being in how many times subjects saw them, or if they saw a still image taken directly from each video.”<sup>99</sup>

Later, the article addressed it again:

“We did not measure audiences’ emotional responses to the stories because emotion has repeatedly been shown not to be a causal agent in moral judgment (Blanchette, Gavigan, & Johnston, 2014; Hauser, 2006).”

“Attribution of responsibility, crime severity, and other factors should be examined in future studies concerned with idiosyncrasies of issues; however, we note again that these or any other differences among the stories do not affect our main finding, which is the ability of still photographs over one-time video to improve moral judgment. We can be assured that still photographs caused higher levels of moral judgment compared with one-time video because that was the only thing that changed. All subjects received all three stories, so any differences due to the stories themselves were controlled because all subjects in the different conditions got all the exact same stories.”<sup>100</sup>

**Meador, Aimee, Lewis Knight, Renita Coleman, and Lee Wilkins. 2015. “Ethics in the Digital Age: A Comparison of the Effects of Moving Images and Photographs on Moral Judgment.” *Journal of Media Ethics* 30 (4): 234–251.**

The most extensive list of threats to internal validity is outlined in Campbell and Stanley’s little book and includes seven different classes of extraneous variables relevant to internal validity in true experiments.<sup>ii</sup> They are briefly described here, along with some practical applications.

### History

This refers to what happens in between the time a treatment is given to a subject and the outcome is measured. For example, if researchers are testing interventions designed to reduce fear of flying and a plane crashes during the study, participants might have more anxiety about flying; the intervention might have worked, but the plane crash

<sup>ii</sup>There is an eighth, which is related to quasi experimental designs. As true experiments are the focus of this book, this chapter reviews the seven related to true experiments.



compromised the validity of the experiment. History also becomes a concern when long-term effects are an important consideration, and the outcome may need to be measured more than once. For example, newly formed attitudes tend to decrease over time, and under some conditions, time lapses actually *increase* persuasive effects.<sup>101</sup> Studies interested in long-lasting persuasion and attitude formation find it important to measure effects after a delay, in which case events out of a researcher's control could threaten the validity of results. The longer the time lapse, the greater the threat.

In other cases, what happens preceding a treatment may influence the outcome. For example, an experiment that looks at how Muslims are framed in manipulated news stories may produce different results if a radicalized individual attacks a nightclub right before the experiment is conducted. Changes in attitudes may be due to the terrorist attack rather than the treatment.

Threats due to history can be controlled with the pretest–posttest control group design, as whatever events might cause a change in the treatment group would produce the same change in the control group. All groups should be run simultaneously so that whatever happens to one, happens to all. In addition, simultaneous sessions help control for differences due to time of day, day of the week, season, and other changes.<sup>102</sup> In many social science disciplines, history is a major threat to a study, so researchers seek to control threats by exposing subjects to the manipulation and then immediately measuring their responses—for example, showing journalists photographs and immediately asking them to solve an ethical dilemma;<sup>103</sup> this fits with what would normally occur in newsrooms on tight deadlines. No delay occurred, so history could not have been a threat to internal validity.

### Maturation

This also is related to the passage of time but is specifically concerned with what happens to the *subject* over time, such as getting older, wiser, hungrier, more tired or bored, etc. Experiments on effective teaching methods can be spread out over weeks, months, semesters, or even a year. During this time, biological and psychological changes can occur in subjects, and that can be the cause of the outcome rather than the treatment. Even a study that takes too long to complete can result in subjects getting bored or fatigued, which may change their responses. Campbell and Stanley give examples of the cumulative effects of learning and pressure of the experience as being the cause of improvements rather than the educational intervention.<sup>104</sup>

### Testing

This refers to the problem discussed earlier about pretests, which are not only a threat to external validity but also to internal validity. When subjects are tested more than

once, they tend to do better or worse, for reasons such as they learned how to take the test or grew bored, among other things. People routinely do better the second time they take personality, achievement, and intelligence tests<sup>105</sup> but show more prejudice on stereotyping tests.<sup>106</sup> All this can occur without any intervention whatsoever. Avoiding the use of pretests unless necessary is one way to circumvent this threat. When a pretest is necessary, using alternate but equivalent forms of the test for pre- and postsessions can help alleviate this concern.<sup>107</sup> Another way to lessen the threat of testing is by using **nonreactive or unobtrusive measures**, which are those that leave subjects unaware that they are being studied, unsuspecting of what is being measured, or unable to alter their responses, such as with measures of sweating, heart rate, or how long they spend looking at something as measured by eye tracking devices or web screen recording software. When subjects are aware that they are being studied and are able to guess the hypothesis, demand characteristics and social desirability come into play. Unobtrusive measures also can be devised so that while subjects know they are being studied, they do not know what the “right” answers are or the specific hypotheses of the study. In moral judgment research, for example, subjects know they are taking an “ethics test” but do not know the “right” answers unless they also know Kohlberg’s moral development theory.<sup>108</sup> Subjects think the important question is what action they would take when the report comes out, not how many statements they choose that represent higher levels of moral judgment. Many more examples of unobtrusive measures will be given in chapter 10. Observing subjects is probably the most popular unobtrusive measure—for example, recording which magazines subjects read while in a waiting room after receiving some treatment. Festinger and Carlsmith’s<sup>109</sup> measure of cognitive dissonance was whether the subjects tried to convince others that the experiment was fun, interesting, and worthwhile, or whether they told the truth that it was boring and worthless. Another observation measure could be whether a subject puts a plastic cup in a recycling bin after receiving some message about the environment.

### Instrumentation

This threat involves issues with the actual instruments, such as galvanic skin response machines that measure the sweat on a subject’s skin, often used to indicate psychological or physiological arousal. When machines break down or need to be recalibrated, this threatens internal validity because it may produce changes in the scores recorded. Another form of instrumentation is human—when people are used to observe behavior, score, or code something, they can vary in their judgment or standards can change. They can also get better with practice or tired over time. Campbell and Stanley<sup>110</sup> recommend randomly assigning observers and using each

one for both experimental and control groups as well as **double blinding** them—that is, not letting either the subjects or the observers know who is getting the treatment so as not to bias the observers' ratings. There are also techniques to determine if standards are different for different scorers, or if standards have changed over the time the scoring is being done, called **interrater reliability**, which is explained in chapter 10. If observed outcomes can be unobtrusively recorded—for example with a video camera or web tracking software—then how similarly the two scorers are measuring things can be assessed.

### Statistical Regression

Also called **regression to the mean**, it represents a threat when extreme scores tend to revert to the mean of the group—for example, when the smartest students appear to be getting duller while the dumbest students appear to be getting smarter. This occurs with the very highest and lowest scores, so researchers are advised not to select subjects for a study *because* they are particularly good or poor at something; they likely will be more average when tested again. So, for example, a teaching intervention that uses only students with the worst test scores will likely show a pseudo effect of the intervention rather than a real effect.<sup>111</sup> When subjects are chosen that represent the full range, the effects of regression to the mean by extreme scores can be controlled by random assignment. Extreme scorers should be equally assigned to both treatment and control groups so that each group regresses to the mean equally.<sup>112</sup>

### Selection

Selection of subjects that results in biases also is controlled by random assignment; whatever differences there are should be equally distributed among the groups.<sup>113</sup> When random assignment is not possible—for example, when students self-select to be in a certain class—the study becomes a quasi experiment, and readers should refer to one of the many good books on this topic for advice on how to lessen this threat.

### Attrition

Also known as mortality, this is when subjects drop out of studies or do not answer all the questions. When the attrition rate is different for treatment and comparison groups, it can compromise validity. When there is systematic attrition—for example, if subjects drop out of the treatment group in higher rates than the control group—then any effects may be due to group differences rather than the treatment. Even if the groups had been equal before the study, dropouts can make them unequal. For example, subjects may drop out of an educational intervention because they do not think it is working for them, or out of an exercise program because they think it is too hard. A poststudy analysis of

group equality will uncover this if it is a problem. Shortening the length of the study can help alleviate attrition concerns.

In summary, the most effective way to control for internal validity is by using randomization and designs with a control group.<sup>114</sup> Randomization includes not only randomly assigning subjects to condition but also randomly assigning as many other elements as possible—for example, randomizing the order of experimental sessions or which experimenters run which session if all groups cannot be tested at the same time. These will be discussed in more depth in chapter 7. Control groups have everything happen to them that the treatment groups have except the treatment; they are subject to the same threats of history, maturation, regression, testing, instrumentation, and attrition. If researchers can control these threats to internal validity, then an experimental design is said to be **rigorous**, meaning the most accurate, precise, or valid that one can design.

Readers will also no doubt realize some contradictions in this chapter—for example, that pretests present a threat to internal validity via testing effects but are one way to alleviate concerns about threats from history. As with all decisions regarding the design of an experiment, researchers must weigh the pros and cons, and carefully consider which issues represent the greatest threat to the validity of a particular experiment, and a well-articulated discussion of the trade-offs should be presented in the paper.

There are other more specific types of internal validity, such as construct validity, face validity, internal consistency and reliability, which will be dealt with in chapter 10. Next, this book turns to more specific decisions to be made on the type of design for an experiment—specifically, how many independent variables will be manipulated and how many each subject will get.

## Common Mistakes

- Designing experiments that are not as realistic as possible
- Not using subjects that are similar to those one intends to make logical inferences about
- Replicating others' findings without extending them in some theoretical way
- Failing to control confounding variables

## Test Your Knowledge

1. Lab experiments are usually weaker in \_\_\_\_\_ validity than in \_\_\_\_\_ validity.
  - a. External, internal
  - b. Internal, external
  - c. Ecological, external
  - d. Logical, statistical
2. Typically, lab experiments use random sampling in order to be generalizable.
  - a. True
  - b. False
3. The idea that reasonable conclusions can be drawn based on common sense without a mathematical calculation of the fit between a sample and a population is referred to as \_\_\_\_\_.
  - a. Statistical inference
  - b. Logical inference
  - c. Nonsensical inference
  - d. Replication
4. Repeating a study to test the theory while also extending it in some theoretically meaningful way is known as \_\_\_\_\_.
  - a. Self-replication
  - b. Direct replication
  - c. External replication
  - d. Conceptual replication
5. Basic criteria for internal validity are \_\_\_\_\_.
  - a. The cause must precede the effect
  - b. The cause and effect must be related
  - c. There are no other plausible explanations for the effect
  - d. All of these
6. Of the seven classes of extraneous variables that threaten internal validity, which one describes when people who are recording observations differ in their judgments or change their standards?
  - a. Selection
  - b. Attrition

For Review-No Commercial Use(2023)

- c. Regression to the mean
  - d. Instrumentation
7. How does random assignment provide internal validity?
    - a. By making samples randomly similar to each other.
    - b. By sampling from the population to be generalized.
    - c. By ensuring that each person or unit from a population has an equal chance of being chosen.
    - d. By ensuring representativeness.
  8. Of the seven classes of extraneous variables that threaten internal validity, which one describes when the smartest students appear to be getting duller, while the dumbest students appear to be getting smarter?
    - a. Selection
    - b. Attrition
    - c. Regression to the mean
    - d. Instrumentation
  9. Of the seven classes of extraneous variables that threaten internal validity, which one describes when subjects drop out of studies or do not answer all the questions?
    - a. Selection
    - b. Attrition
    - c. Regression to the mean
    - d. Instrumentation
  10. Of the seven classes of extraneous variables that threaten internal validity, which one describes when students choose to be in a certain class?
    - a. Selection
    - b. Attrition
    - c. Regression to the mean
    - d. Instrumentation

### Answers

- |      |      |      |       |
|------|------|------|-------|
| 1. a | 4. d | 7. a | 9. b  |
| 2. b | 5. d | 8. c | 10. a |
| 3. b | 6. d |      |       |

## Application Exercises

1. Continue to develop your experiment by writing about how you will ensure external and internal validity in your study. Write one or two pages on how you will increase external validity by making your study more realistic and generalizable. Use the topics in this chapter as a guide. Write one or two more pages on how you will minimize threats to internal validity, specifically considering the seven threats described in the chapter. Also include a discussion of the trade-offs, as you can rarely have perfect internal and external validity. Which is more important, and what would you give up; why? Imagine reviewers have challenged you on these issues and respond to their concerns.
2. Find three published experiments in your field and write a one-page critique of each on how the authors addressed issues of internal and external validity. What trade-offs did they make? Was it convincing? Are there other arguments you would have used? Are there other threats they did not consider but should have?

## Suggested Readings

Chapter 2, “Statistical Conclusion Validity and Internal Validity,” and part of chapter 3, “External Validity,” pp. 83–102, in Shadish, W. R., F. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth. You do not need to read the Construct Validity section yet; that will come in the next chapter.

For a good example of experimental realism, read Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American opinion*. Chicago: University of Chicago Press. These studies simulate actual national news broadcasts as much as possible.

Stoker, Gerry. 2010. “Exploring the Promise of Experimentation in Political Science: Micro-Foundational Insights and Policy Relevance.” *Political Studies* 58: 300–319. This paper discusses issues of validity in arguing for more experimentation in political science, but the arguments are applicable to all of social science.

## Notes

1. John A. Courtright, “Rationally Thinking About Nonprobability,” *Journal of Broadcasting & Electronic Media* 40, no. 3 (1996): 414.
2. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. (Chicago: Rand McNally, 1963).
3. E. A. Stuart et al., “Estimates of External Validity Bias When Impact Evaluations Select Sites Purposively.” Paper presented at the Proceedings from the Society for Research on Annual Effectiveness (Washington, DC, Spring 2012).
4. Rebecca B. Morton and Kenneth C. Williams, *Experimental Political Science and the Study of*

- Causality: From Nature to the Lab* (New York: Cambridge University Press, 2010).
5. Murray Webster and Jane Sell, *Laboratory Experiments in the Social Sciences* (Amsterdam: Elsevier, 2007).
  6. Morton and Williams, *Experimental Political Science*.
  7. Maria Jimenez-Buedo and Francesco Guala, "Artificiality, Reactivity, and Demand Effects in Experimental Economics," *Philosophy of the Social Sciences* 46, no. 1 (2016): 3–23.
  8. Renita Coleman and Esther Thorson, "The Effects of News Stories That Put Crime and Violence into Context: Testing the Public Health Model of Reporting," *Journal of Health Communication* 7, no. 4 (2002): 401–426.
  9. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*, 17.
  10. Morton and Williams, *Experimental Political Science*; Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  11. Morton and Williams, *Experimental Political Science*, 253.
  12. Elizabeth Tipton, "How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations," *Journal of Educational and Behavioral Statistics* 39, no. 6 (2014): 478–501.
  13. Roger Kirk, *Experimental Design: Procedures for the Behavioral Sciences*, 4th ed. (Thousand Oaks, CA: Sage, 2013).
  14. Mirae Kim and Gregg G. Van Ryzin, "Impact of Government Funding on Donations to Arts Organizations: A Survey Experiment," *Nonprofit and Voluntary Sector Quarterly* 43, no. 5 (2014): 910–925.
  15. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  16. Morton and Williams, *Experimental Political Science*.
  17. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*, 19.
  18. Ibid.
  19. Annie Lang, "The Logic of Using Inferential Statistics with Experimental Data From Nonprobability Samples: Inspired by Cooper, Dupagne, Potter, and Sparks," *Journal of Broadcasting and Electronic Media* 40, no. 3 (Summer 1996): 422–430; Courtright, "Rationally Thinking About Nonprobability"; Michael D. Basil, "The Use of Student Samples in Communication Research," *Journal of Broadcasting and Electronic Media* 40, no. 3 (Summer 1996): 431–440.
  20. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  21. Tipton, "How Generalizable Is Your Experiment?"
  22. For example, see Elizabeth Tipton, "Stratified Sampling Using Cluster Analysis: A Balanced-Sampling Strategy for Improved Generalizations from Experiments," *Evaluation Review* 37 (2014): 109–139; E. A. Stuart, "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science* 25 (2010): 1–21; Stuart et al., "Estimates of External Validity Bias."
  23. L. V. Hedges and C. A. O'Muircheartaigh, "Improving Generalizations from Designed Experiments," (Evanston, IL: Northwestern University, 2011). Quoted in Tipton, 2014.
  24. Tipton, "How Generalizable Is Your Experiment?"
  25. David Freedman, Robert Pisani, and Roger Purves, *Statistics*, 4th ed. (New York: Norton, 2007), 335.
  26. William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Belmont, CA: Wadsworth Cengage Learning, 2002), 342.
  27. Ibid., 248.
  28. For example, see Edmund J. Malesky, Dimitar D. Gueorguiev, and Nathan M. Jensen, "Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment," *American Journal of Political Science* 59, no. 2 (2015): 419–439.
  29. Lang, "The Logic of Using Inferential Statistics."
  30. Malesky, Gueorguiev, and Jensen, "Monopoly Money."
  31. Ibid., 426–427.
  32. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 373.
  33. Ibid., 23.
  34. Ibid., 348.
  35. P. Rossi, M. Lipsey, and H. Freeman, *Evaluation: A Systematic Approach*, 7th ed. (Thousand Oaks, CA: Sage, 2004).
  36. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 355.
  37. Eric Kramon, "Where Is Vote Buying Effective? Evidence from a List Experiment in Kenya," *Electoral Studies* 44 (2016): 397–408.



38. For examples, see Arnstein Øvrum and Elling Bere, "Evaluating Free School Fruit: Results from a Natural Experiment in Norway with Representative Data," *Public Health Nutrition* 17, no. 6 (2014): 1224–1231.
39. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
40. Kimberly Klaiman, David L. Ortega, and Cloé Garnache, "Perceived Barriers to Food Packaging Recycling: Evidence from a Choice Experiment of US Consumers," *Food Control* 73 (2017): 291–299.
41. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
42. Arjan Verschoor, Ben D'Exelle, and Borja Perez-Viana, "Lab and Life: Does Risky Choice Behaviour Observed in Experiments Reflect That in the Real World?" *Journal of Economic Behavior & Organization* 128 (2016): 134–148.
43. Ibid., 136.
44. Alese Wooditch, Lincoln B. Sloas, and Faye S. Taxman, "A Multisite Randomized Block Experiment on the Seamless System of Care Model for Drug-Involved Probationers," *Journal of Drug Issues* 47, no. 1 (2017): 50–76.
45. For example, see Malcolm Fairbrother, "Geoengineering, Moral Hazard, and Trust in Climate Science: Evidence from a Survey Experiment in Britain," *Climatic Change* 139, no. 3/4 (2016): 477–489; Mogens Jin Pedersen and Justin M. Stritch, "Internal Management and Perceived Managerial Trustworthiness," *The American Review of Public Administration* (2016): 1–20.
46. For example, see Øvrum and Bere, "Evaluating Free School Fruit"; Rosie Campbell and Philip Cowley, "What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment," *Political Studies* 62, no. 4 (2013): 745–765; Tianguang Meng, Jennifer Pan, and Ping Yang, "Conditional Receptivity to Citizen Participation," *Comparative Political Studies* (2014): 1–35; Klaiman, Ortega, and Garnache, "Perceived Barriers to Food Packaging"; M. L. Loureiro and W. J. Umberger, "A Choice Experiment Model for Beef: What U.S. Consumer Responses Tell Us About Relative Preferences for Food Safety, Country-of-Origin Labeling and Traceability," *Food Policy* 32, no. 4 (2007): 496–514.
47. For examples, see Claire A. Dunlop and Claudio M. Radaelli, "Teaching Regulatory Humility: Experimenting with Student Practitioners," *Politics* 36, no. 1 (2016): 79–94; Jin Huang et al., "Individual Development Accounts and Homeownership Among Low-Income Adults With Disabilities: Evidence From a Randomized Experiment," *Journal of Applied Social Science* 10, no. 1 (2016): 55–66; Wooditch, Sloas, and Taxman, "A Multisite Randomized Block Experiment."
48. Lang, "The Logic of Using Inferential Statistics."
49. Ibid.
50. Courtright, "Rationally Thinking About Nonprobability."
51. Lang, "The Logic of Using Inferential Statistics."
52. Ibid.
53. Basil, "The Use of Student Samples."
54. Joanne M. Miller, "Examining the Mediators of Agenda Setting: A New Experimental Paradigm Reveals the Role of Emotions," *Political Psychology* 28, no. 6 (2007): 689–717.
55. Basil, "The Use of Student Samples."
56. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
57. C. Cook, M. Umberger, and T. J. Landrum, "Determining Evidence-Based Practices in Special Education," *Exceptional Children* 75 (2009): 365–383.
58. S. Schmidt, "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences," *Review of General Psychology* 13 (2009): 7.
59. Matthew C. Makel et al., "Replication of Special Education Research: Necessary But Far Too Rare," *Remedial and Special Education* 37, no. 3 (2016): 1–8; *ibid.*
60. Gary King, "Replication, Replication," *PS: Political Science and Politics* 28, no. 4 (1995): 444–452.
61. Susan T. Fiske, "How to Publish Rigorous Experiments in the 21st Century," *Journal of Experimental Social Psychology* 66 (2016): 145–147.
62. Makel et al., "Replication of Special Education Research."
63. Matthew C. Makel and Jonathan A. Plucker, "Facts Are More Important Than Novelty: Replication in the Education Sciences," *Educational Researcher* 43 (2014): 304–316.
64. Matthew C. Makel, Jonathan A. Plucker, and B. Hegarty, "Replications in Psychology Research: How Often

- Do They Really Occur?" *Perspectives on Psychological Science* 7 (2012): 537–542.
65. S. A. Haslam and C. McGarty, "100 Years of Certitude? School Psychology, the Experimental Method, and the Management of Scientific Uncertainty," *British Journal of Social Psychology* 40 (2001): 1–21.
  66. Renita Coleman, "The Effect of Visuals on Ethical Reasoning: What's a Photograph Worth to Journalists Making Moral Decisions?" *Journalism and Mass Communication Quarterly* 83, no. 4 (Winter 2006): 835–850. I use examples from my own work not because they are paragons of excellence in experimental design but because I know the backstory and reasoning process behind them and can explain it from a firsthand perspective.
  67. Renita Coleman, "Race and Ethical Reasoning: The Importance of Race to Journalistic Decision Making," *Journalism and Mass Communication Quarterly* 80, no. 2 (2003): 295–310; Renita Coleman, "Color Blind: Race and the Ethical Reasoning of African Americans on Journalism Dilemmas," *Journalism and Mass Communication Quarterly*, 88, no. 2 (Summer 2011): 337–350; Renita Coleman, "The Moral Judgment of Minority Journalists: Evidence from Asian American, Black, and Hispanic Professional Journalists," *Mass Communication and Society* 14, no. 5 (September–October 2011): 578–599.
  68. Aimee Meader et al., "Ethics in the Digital Age: A Comparison of the Effects of Moving Images and photographs on Moral Judgment," *Journal of Media Ethics* 30, no. 4 (2015): 234–251.
  69. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 342; Makel et al., "Replication of Special Education Research"; Fiske, "How to Publish Rigorous Experiments."
  70. Christian S. Crandall and Jeffrey W. Sherman, "On the Scientific Superiority of Conceptual Replications for Scientific Progress," *Journal of Experimental Social Psychology* 66 (2016): 93–99.
  71. Ibid.; Makel et al., "Replication of Special Education Research"; Makel et al., "Facts Are More Important Than Novelty"; Makel et al., "Replications in Psychology Research."
  72. B. D. McCullough, Kerry Anne McGeary, and Teresa D. Harrison, "Lessons from the JMCB Archive," *Journal of Money, Credit, and Banking* 38, no. 4 (2006): 1093–1107.
  73. Crandall and Sherman, "On the Scientific Superiority."
  74. Stanley Milgram, *Obedience to Authority: An Experimental View* (New York: Harper & Row, 1974).
  75. Fiske, "How to Publish Rigorous Experiments."
  76. Esther Thorson, Robert H. Wicks, and Glenn Leshner, "Experimental Methodology in Journalism and Mass Communication Research," *Journalism and Mass Communication Quarterly* 89, no. 1 (2012): 112–124.
  77. Fiske, "How to Publish Rigorous Experiments."
  78. Tim Schweizer, "Introducing Idea Mapping into the Business Curriculum: Results from Two Experiments," *The International Journal of Technology, Knowledge and Society* 7, no. 1 (2011): 190–200.
  79. Pazit Ben-Nun Bloom, "Disgust, Harm and Morality in Politics," *Political Psychology* 35, no. 4 (2014): 495–513.
  80. Makel et al., "Replication of Special Education Research"; Makel et al., "Facts Are More Important Than Novelty"; Makel et al., "Replications in Psychology Research."
  81. Ibid.
  82. Crandall and Sherman, "On the Scientific Superiority."
  83. Charles R. Ebersole et al., "Observe, Hypothesize, Test, Repeat: Luttrell, Petty and Xu (2017) Demonstrate Good Science," *Journal of Experimental Social Psychology* 69 (2017): 184–186.
  84. J. P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLoS Medicine* 2 (2005): 696–701; Paul H. P. Hanel and Katia C. Vione, "Do Student Samples Provide an Accurate Estimate of the General Public?" *PLoS ONE* 11, no. 12 (2016): 1–10.
  85. Rebecca S. McEntee, Renita Coleman, and Carolyn Yaschur, "Comparing the Effects of Vivid Writing and Photographs on Moral Judgment in Public Relations," *Journalism and Mass Communication Quarterly* 94 no. 4 (2017): 1011–1030.
  86. Fiske, "How to Publish Rigorous Experiments."
  87. McCullough, McGeary, and Harrison, "Lessons from the JMCB Archive."
  88. Eric M. Camburn et al., "Benefits, Limitations, and Challenges of Conducting Randomized Experiments

- with Principals,” *Educational Administration Quarterly* 52, no. 2 (2016): 187–220; J. A. Maxwell, “Causal Explanation, Qualitative Research, and Scientific Inquiry in Education,” *Educational Researcher* 33, no. 2 (2004): 3–11.
89. J. K. Maner, “Into the Wild: Field Studies Can Increase Both Replicability and Real-World Impact,” *Journal of Experimental Social Psychology* 66 (2016): 100–106; Fiske, “How to Publish Rigorous Experiments”; Camburn et al., “Benefits, Limitations, and Challenges.”
  90. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  91. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*.
  92. Thorson, Wicks, and Leshner, “Experimental Methodology in Journalism.”
  93. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*, 5.
  94. Cengiz Erisen, Elif Erisen, and Binnur Ozkececi-Taner, “Research Methods in Political Psychology,” *Turkish Studies* 13, no. 1 (2013): 13–33.
  95. Gustav Fischer, “Empirische Sozialforschung,” *Jahrgang* 44, no. 2 (1936).
  96. Coleman, “Race and Ethical Reasoning”; “Color Blind: Race and the Ethical Reasoning of African Americans on Journalism Dilemmas”; “The Moral Judgment of Minority Journalists: Evidence from Asian American, Black, and Hispanic Professional Journalists.”
  97. Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth, *Emotion in the Human Face* (New York: Pergamon, 1972); A. Mehrabian, “Inference of Attitudes from the Posture, Orientation, and Distance of a Communicator,” *Journal of Consulting and Clinical Psychology* 32, no. 3 (June 1968): 296–308; “Some Referents and Measures of Nonverbal Behavior,” *Behavioral Research Methods and Instrumentation* 1 (1969): 213–217; *Silent Messages* (Belmont, CA: Wadsworth, 1971); *Nonverbal Communication* (Chicago: Aldine/Atherton, 1972); J. K. Burgoon, T. Birk, and M. Pfau, “Nonverbal Behaviors, Persuasion, and Credibility,” *Human Communication Research* 17, no. 1 (Fall 1990): 140–169.
  98. Thorson, Wicks, and Leshner, “Experimental Methodology in Journalism.”
  99. Meader et al., “Ethics in the Digital Age,” 245.
  100. Ibid., 247.
  101. Alice H. Eagly and Shelly Chaiken, *The Psychology of Attitudes* (Fort Worth, TX: Harcourt Brace Jovanovich, 1993).
  102. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  103. Coleman, “The Effect of Visuals on Ethical Reasoning.”
  104. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*, 8–9.
  105. Anne Anastasi, *Differential Psychology*, 3rd ed. (New York: Macmillan, 1958); V. R. Cane and A. W. Heim, “The Effects of Repeated Testing: III. Further Experiments and General Conclusions,” *Quarterly Journal of Experimental Psychology* 2 (1950): 182–195; C. Windle, “Test-Retest Effect on Personality Questionnaires,” *Educational Psychology Measurement* 14 (1954): 617–633.
  106. R. E. Rankin and D. T. Campbell, “Galvanic Skin Responses to Negro and White Experimenters,” *Journal of Abnormal and Social Psychology* 51 (1955): 30–33.
  107. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  108. Lawrence Kohlberg, *The Philosophy of Moral Development: Moral Stages and the Idea of Justice* (Cambridge, MA: Harper & Row, 1981); *The Psychology of Moral Development: The Nature and Validity of Moral Stages* (San Francisco: Harper & Row, 1984).
  109. L. Festinger and J. M. Carlsmith, “Cognitive Consequences of Forced Compliance,” *Journal of Abnormal and Social Psychology* 58, no. 2 (1959): 203–210.
  110. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  111. Ibid.
  112. Ibid.
  113. Ibid.
  114. Ibid.



For Review-No Commercial Use(2023)



# FACTORIAL DESIGNS

*Needless complexity seldom makes for better experimental research.<sup>1</sup>*

—Diana Mutz

## LEARNING OBJECTIVES

- Identify single-factor and factorial designs.
- Diagram an experiment using factorial notation and design tables.
- Explain the differences among between- and within-subjects designs, mixed and incomplete factorials.
- Create an experiment and explain the use of factors and how subjects are assigned.
- Plan how to implement a control group.

In chapter 4, this text discussed the different kinds of experiments in broad strokes, including natural and field experiments, true and quasi experiments, and the typologies such as the pretest–posttest control group design and Solomon four-group design. This chapter hones in on one type of experiment—the true or laboratory experiment—and drills down into more specific decisions about designing it. Specifically, this chapter looks at the number of factors in an experiment, how subjects are used in different designs, and control groups. When reading this chapter, keep in mind the advice in the opening quote, as things can get complicated quickly.

## SINGLE-FACTOR DESIGNS

---

*Factor* is another word for the independent variable that is manipulated—that is, the treatment of interest, or the thing deliberately changed in a true experiment. So far, this book has stressed the importance of manipulating one thing that is the independent variable of interest, whatever is hypothesized to cause a change in some outcome or have an effect on some dependent variable. When there is only one factor in an experiment, it is called a **single-factor, one-factor, or one-way design**.<sup>2</sup> For example, in a study of news stories, the factor to be manipulated might be how stories are framed or the angle that they use. In an interpersonal communication study, the factor of interest might be lying. In a psychology study, researchers could be interested in people's feelings of insecurity. In a political science experiment, the factor may be political position. Framing, lying, political position, and insecurity are the theoretical concepts considered to be the factors.

For any factor, there must be two or more **levels**, or different discrete values of each.<sup>3</sup> In a study of news frames, one researcher<sup>4</sup> studied two levels: objective and advocacy frames. He defined advocacy frames as one-sided stories that gave the perception of consensus, and objective frames as stories written as two-sided, detached news stories that did not give a perception of consensus. In a study of insecurity, researchers used two levels of the factor—high or none—by having subjects either write about their financial futures (creating high insecurity) or write about listening to music, a neutral topic not expected to create any feelings of insecurity.<sup>5</sup> A political science study used two levels of political position by labeling one position as liberal and another as conservative.<sup>6</sup>

Three and even four levels of a factor are also common. For instance, in a study of lying, researchers used three levels: lying by omission, truthful statements that were misleading, and the truth.<sup>7</sup> It is unusual to see more than this because too many can make a study unwieldy. It is more common to conduct another experiment if more than a few factors with many levels are investigated.

## FACTORIAL DESIGNS

---

The single-factor experiment is one of the simplest designs and is a good way to begin learning to do experiments. In reality, however, most experiments manipulate more than one independent variable at a time, making it a **factorial design**. Factorial designs look at the effect of more than one variable at the same time but still independently of each other. This is a more efficient method of testing hypotheses, as two or more things can be studied at one time instead of having to conduct multiple separate experiments.<sup>8</sup> Most experiments were single-factor designs up until Ronald Fisher started arguing for multiple factors being

more economical, coining the term *factorial* in his book *Design of Experiments*.<sup>9</sup> Campbell and Stanley call the one-variable-at-a-time approach ineffective and a “blunt tool.”<sup>10</sup> The value added by a factorial design is that it not only allows researchers to see what happens when each of the factors is changed independently but also allows them to see what happens when the factors are combined. This ability to look at how phenomena depend on the joint actions of two or more variables represents the complex world more realistically, as many outcomes are produced by several factors at a time.<sup>11</sup>

For example, when people see a public service announcement (PSA) on TV, they typically are influenced by a variety of characteristics of the PSA. In a study on PSAs about HIV/AIDS, researchers examined three of those things, represented by the following factors: (1) the quality of the argument in the ad, with the levels being a strong or weak argument; (2) how personally relevant the subject of HIV/AIDS was to each subject (high/low); and (3) two different types of ad formats, one that used a narrative storyline and one that relied on statistical data.<sup>12</sup> The results showed that strong arguments were significantly better than weak ones in changing people's attitudes toward and intentions to use condoms, but that ad format did not matter; the narrative and statistical formats worked about the same. Also, subjects who said HIV/AIDS was more relevant to them personally were more likely to intend to use condoms. These three findings are the **main effects**—that is, the ability for the different levels in one factor alone to cause a change in the outcome variables.<sup>13</sup> It represents the difference between *levels* of one factor and has nothing to say about any other factor. In the prior example, strong arguments produced changes in subjects' attitudes and intention when weak arguments did not; thus, there was a main effect of argument strength. There was also a main effect of personal relevance, with high relevance being more effective than low relevance. But there was no main effect of ad format; regardless of whether the ad was in narrative or statistical format, the effect on the outcome was the same.

The study in this example used three factors, also called a three-way design. Experiments with two factors are called two-way designs and are the simplest version of a factorial experiment, containing two factors with two levels each. Terms for the number of factors in an experiment are intuitive—four-way for four factors, etc. However, four-way designs and higher are rare and more complicated to analyze and interpret, so they are not found as often.

## Main Effects and Interactions

Several outcomes are possible with a factorial design: the main effect of each of the factors by itself, described previously, and the effect of the two or more factors in



tandem—that is, if the effect of one factor *changes* depending on the other.<sup>14</sup> When the outcome of one factor depends on another, this is known as an **interaction effect**. In other words, the effects of two factors acting in combination could not be predicted from knowing the effects of each one separately.<sup>15</sup> An interaction is between factors, not levels, the way a main effect is. An easy way to think about this is with diet and exercise being two factors that affect your weight. If changing your diet alone affects your weight, that is a main effect; if changing your exercise program alone affects your weight, that is a main effect. When both diet and exercise are changed at the same time, there can be an interaction—for example, if you only lose weight when you eat a low-fat diet and also exercise for three more days a week. In the study of HIV/AIDS messages described earlier, there was an interaction effect where strong arguments and low personal relevance together had an effect. To report an interaction, you first say which of the factors interacted with each other and then go on to describe the levels of each factor that produced the effect. For example, the HIV/AIDS study would say something like, “The factors of argument strength and personal relevance interacted. For subjects who found HIV/AIDS to be of little personal relevance, strong arguments were significantly better at encouraging them to use condoms.”<sup>16</sup> This study shows the benefits of studying three factors in one experiment; the researchers would not have known that strong arguments work for people with low personal relevance had they studied argument strength in one study and relevance in another.

A factorial design can result in different combinations of outcomes:

- Significant main effects for all the factors but no interaction effect
- Significant main effects for one or some of the factors but not all of them, with either a significant or nonsignificant interaction
- A significant interaction but no main effects<sup>17</sup>

Campbell and Stanley<sup>18</sup> give an example of a hypothetical study of three different teacher personality types as one factor, and three different teaching methods as the other factor. If all three personalities and methods are equally effective, there would be no main effects for any of them; that is, none of the different levels of the factors are significantly better than any others within that factor. But the teaching method and teacher personalities could interact so that, for example, the spontaneous teachers do better with group discussion than with tutorials.

In many studies, the interaction effect is really what researchers are interested in. For example, in a health communication study, researchers point out that framing messages



differently has little effect,<sup>19</sup> saying that this may be because people's assessments of risk need to be considered along with framing effects. This study is specifically designed to examine the interaction of framing and risk assessment, and is less interested in the main effects of either factor, primarily because the main effects are already known.<sup>20</sup>

When conducting a factorial design, hypotheses should be predicted for each factor independently (the main effects), and either a hypothesis or a research question (RQ) posed for each of the possible interactions. For example, a study with two factors such as the teaching method and teacher personality study will have two hypotheses that predict main effects—one for teacher personality and one for teaching methods. It will also have a third hypothesis or RQ about the interaction between the two factors—that teacher personality and teaching method will interact. If there is not enough evidence to make a prediction for possible interaction effects, researchers typically ask an RQ; here is an example of one phrased in narrative instead of a formal RQ. It is from a study that looked at whether different types of warnings could prevent people from buying accessories like wireless speakers that were incompatible with their electronic devices like cell phones; in this field, warning messages are called a “nudge”:

### For Review-No Commercial Use(2023)

The study also tested for interaction effects between the socio-demographic variables (age and education) and the experimental treatments (warning message . . .). In other words, if there were an effect of age and education on the purchase of incompatible products, would this effect be the same in all experimental conditions? Or could it be that this effect is increased (or reduced) in the presence of a given type of nudge? Since it is difficult to predict the direction and magnitude of these interaction effects, this part of the investigation was exploratory and not guided by specific hypotheses.<sup>21</sup>

Alternately, a formal RQ could be written:

RQ: Is there an interaction between sociodemographics (age and education) and warning message style (traditional, emotive, no warning)?

In another example, using a study of framing health messages and the level of risk people had for contracting the human papillomavirus (HPV), researchers were mainly interested in how risk and framing depended on each other, so they proposed two hypotheses that only predicted interactions; there were no predictions of main effects.<sup>22</sup> They also had enough evidence to make predictions of direction for the interaction effects. The italicized words (mine) represent the level of each factor predicted to interact.

“H1: When parents perceive their children to be at *low risk* of contracting HPV infections, a *loss-framed* (vs. gain-framed) message will induce greater intentions to vaccinate their children (a) free of cost and (b) with cost.”

“H2: When parents perceive their children to be at *high risk* of contracting HPV infections, a *gain-framed* (vs. loss-framed) message will induce greater intentions to vaccinate their children (a) free of cost and (b) with cost.”<sup>23</sup>

To put it another way, low-risk people are expected to have greater intentions to vaccinate when they see loss-framed messages, and high-risk people when they see gain-framed messages.

I recommend not getting carried away with too many factors or too many levels of each factor, because the experiment can become so complex that it gets unwieldy on a number of fronts. For example, the number of interactions goes up exponentially with each factor. In an experiment with two factors, there is one possible interaction (Factor 1 with 2). But add a third factor and there is a possibility of four interactions (Factor 1 with 2; Factor 1 with 3; Factor 2 with 3; and between all three, Factors 1, 2, and 3). Experiments with ten factors at two levels produces  $2^{10} = 1,024$  combinations! And it can get tricky to interpret the meaning of interactions where there are more than two factors.<sup>24</sup> It is better to control all other variables that might affect the outcome and manipulate them in different studies.

In addition to the number of interactions, designs with multiple factors and multiple levels of each require creating more **stimuli**, which are the way treatments are conveyed—for example, through messages, advertisements, or teaching technique. In one political communication study, the researcher had to write 256 press releases because the study used five factors with two to four levels for each.<sup>25</sup> Fortunately, the press releases were short.

Not only are the number of stimuli an issue, the number of subjects needed also increases, which will be covered in a later chapter. Interactions may call for larger sample sizes because they are harder to detect than main effects.<sup>26</sup> Factorial designs can require fewer subjects than separate single-factor studies if each subject is exposed to both factors,<sup>27</sup> but the benefits decrease at a certain point because of the need to perform multiple statistical tests, making it harder to find significance unless even more subjects are added.<sup>28</sup> In short, it can be better to design several sequential experiments instead of trying to examine too much in one study. Two, three, or four factors are the most commonly used factorial designs.<sup>29</sup>

## Factorial Notation

When a factorial design is used, it is standard practice to refer to it using **factorial notation**—for example, as a 2 x 2 design, a 3 x 3 design, or a 2 x 2 x 3, with the “x” pronounced “by” as in “two by two.” Each number refers to a factor, so the number of

MORE ABOUT . . . BOX 6.1

Factorial Notation

In factorial notation, the number of numbers tells how many factors or variables are in the experiment. The value of the numbers tells how many levels of the factor there are. Table 6.1 explains some of the most common factorial designs.

TABLE 6.1 ■ COMMON FACTORIAL DESIGNS

Design notation	The number of numbers tells how many factors	The first number tells how many levels of the first factor	The second number tells how many levels of the second factor	The third number tells how many levels of the third factor	Multiplying gives the number of cells
2 x 2	2 numbers = 2 factors or IVs	The first factor/IV has 2 levels	The second factor/IV has 2 levels	N/A—there is no 3rd number	2 x 2 = 4 cells
2 x 3	2 numbers = 2 factors or IVs	The first factor/IV has 2 levels	The second factor/IV has 3 levels	N/A—there is no 3rd number	2 x 3 = 6 cells
3 x 3	2 numbers = 2 factors or IVs	The first factor/IV has 3 levels	The second factor/IV has 3 levels	N/A—there is no 3rd number	3 x 3 = 9 cells
2 x 2 x 2	3 numbers = 3 factors or IVs	The first factor/IV has 2 levels	The second factor/IV has 2 levels	The third factor/IV has 2 levels	2 x 2 x 2 = 8 cells
2 x 2 x 3	3 numbers = 3 factors or IVs	The first factor/IV has 2 levels	The second factor/IV has 2 levels	The third factor/IV has 3 levels	2 x 2 x 3 = 12

numbers tells how many factors there are. The value of each number refers to the number of levels of that factor. A 2 x 2 means there are two factors, and each factor has two levels. A 3 x 3 means there are two factors because there are only two numbers. Each of the two factors has three levels. A 2 x 2 x 3 means there are three factors, as indicated by the presence of three numbers, with the first and second factor having two levels each and the third factor having three levels. (See More About box 6.1, Factorial Notation.)

I know of no rules about the order in which the factors are represented in the notation; some researchers put the factors with more levels first, and others put the most important factors first. The beginning of an experiment's methods section should describe the design using this notation, usually incorporating the names of the factors and the levels in parentheses, like this:

- “We employed a 2 (argument quality: strong/weak) x 2 (personal relevance: high/low) x 2 (evidence: narrative/statistical) mixed factorial design.”<sup>30</sup>
- “The second study used a 2 (crisis stage: during the crisis vs. postcrisis) x 2 (visual nonverbal cues: powerless vs. powerful) between-subjects factorial design.”<sup>31</sup>

A single factor design can be reported like this:

- “Study 1 used a single-factor design in order to examine the impact of lowered and raised voice pitch on the perceptions of an organizational spokesperson.”<sup>32</sup> In this example, the two levels of the single factor “voice pitch” are raised and lowered.

For more examples of how to report factorial designs, see How To Do It box 6.2.

## For Review-No Commercial Use(2023)

### HOW TO DO IT 6.2

#### Reporting Factorial Design

The portion of the methods section that describes the factorial design, levels of each, and the control group comes at the beginning of the methods section and is usually very brief, ranging from one or two sentences to a paragraph. It should include all of the following:

- Whether it is a single-factor or factorial design
- The number of factors and levels of each using factorial notation. If it is an incomplete factorial, that should be noted here.
- A short description of each level contained within parentheses immediately after the notation
- A description of how subjects are assigned—between-subjects, within-subjects, or as a mixed design
- Whether it has a control group or not and if so, what it is
- Also note the use of random assignment, the subject of the next chapter.

Not all studies follow this exact approach, with some describing these features in a more narrative style. However, it is always safe to do it this way.

Next are examples of how to describe the design, adapted from real studies in order to ensure that they include all of the previous points.

### Single-Factor Design 1

This was a single-factor, between-subjects experiment reflecting the five levels of internal management reviewed previously (goal, commitment, participation, feedback, reward). Each of the subjects was randomly assigned to one of five treatment groups or a control condition. Respondents in the treatment groups received the exact same information as those in the control groups. In addition, subjects in the treatment group were provided one line of text describing the managerial activities.

Adapted from: Pedersen, Mogens Jin, and Justin M. Stritch. 2016. "Internal Management and Perceived Managerial Trustworthiness: Evidence From a Survey Experiment." *The American Review of Public Administration* 48 (1): 67–81. <https://doi.org/10.1177%2F0275074016657179>.

### Single-Factor Design 2

We use a single-factor, between-subjects experiment with subjects randomly assigned to either two levels of the treatment of source of suggestions for citizen participation (formal institutions/Internet) or a control group. The control group received identical information with the exception of a treatment item concerning suggestions of residents from either formal institutions or the Internet.

Adapted from: Meng, Tianguang, Jennifer Pan, and Ping Yang. 2014. "Conditional Receptivity to Citizen Participation: Evidence From a Survey Experiment in China." *Comparative Political Studies* 50 (4): 399–433. <https://doi.org/10.1177%2F0010414014556212>.

### Between-Subjects Factorial 1

This experiment had a 3 (reminder frequency: no reminder, monthly, or weekly reminders) x 2 (method of donation: standing order or one-off donation) factorial design resulting in a total of six treatments. Subjects were randomly assigned to one of the six conditions.

(Note: This study did not specify a control group, but the no-reminder condition represents a baseline for that factor.)

Adapted from: Sonntag, Axel, and Daniel John Zizzo. 2015. "On Reminder Effects, Drop-Outs and Dominance: Evidence from an Online Experiment on Charitable Giving." *PLoS ONE* 10, no. 8 (August 7): 1–17.

### Between-Subjects Factorial 2

This study was a 2 x 3 between-subjects design with the first factor being school level of achievement (High vs. Low) with subjects randomly assigned to be instructed with valid experimental designs only (Valid) or with a mixture of confounded and valid designs (Invalid), and teaching only after the posttest (Control condition).

Adapted from: Lorch, Robert F. Jr., Elizabeth P. Lorch, Benjamin Dunhan Freer, Emily E. Dunlap, Emily C. Hodell, and William J. Calderhead. 2014. "Using Valid and Invalid Experimental Designs to Teach the Control of Variables Strategy in Higher and Lower Achieving Classrooms." *Journal of Educational Psychology* 106, no. 1 (February): 18–35.

(Continued)

(Continued)

### Within-Subjects Factorial

There were three within-subjects factors, each with two levels. The gaze cue factor manipulated the cue face's gaze direction; in the cued condition, the cue face looked toward the target face, while in the uncued condition the cue face looked away from the target face, toward the empty side of the screen. The emotion factor was the manipulation of the cue face's emotional expression (either positive or negative). The number of cues factor was the single or multiple cue face manipulation. There was one cue face in the single cue face condition. All three cue faces were presented in the multiple cue face condition.

(Note: Obviously, when all conditions are within-subjects, participants are assigned to all groups, and designating that they were "randomly assigned" is unnecessary. Also, no control group is designated, as this is one of those situations where a face cannot have no gaze direction. In this study, the groups are compared to each other rather than to a control group that got no treatment.)

From: Landes, Todd Larson, Yoshihisa Kashima, and Piers D. L. Howe. 2016. "Investigating the Effect of Gaze Cues and Emotional Expressions on the Affective Evaluations of Unfamiliar Faces." *PLoS ONE* 11, no. 9: 1–24.

### Mixed Factorial

This was a 4 x 2 x 2 mixed factorial design. The first within-subjects factor was issue (prostitution, drugs, elder abuse, domestic violence). All participants received all four dilemmas. The second within-subjects factor was decision type (with the photograph, use anonymous and real); all participants received two dilemmas that asked whether they should run a photograph after someone associated with the subject requested it not run, and two dilemmas that asked whether a story should run when there are many sources but none who would allow their names to be used. The between-subjects factor was photograph condition (with photograph/without photograph). Subjects were randomly assigned to either the treatment group, where they saw photographs with all four dilemmas, or the control group, where they did not see photographs.

*Adapted from:* Coleman, Renita. 2006. "The Effect of Visuals on Ethical Reasoning: What's a Photograph Worth to Journalists Making Moral Decisions?" *Journalism and Mass Communication Quarterly* 83, no. 4 (Winter): 835–50.

## Design Tables

Another way to represent a factorial design is with a **design table**. These are used to help in planning an experiment and do not seem to be reported in journal articles very often. Tables are also used to determine the characteristics of the stimuli and the number of subjects in each condition, represented by the "cells," or individual boxes. Table 6.2 represents a hypothetical 2 x 2 factorial design in which one factor is whether a person in a photograph is smiling or not, and the other is whether the camera distance is close-up or far away in what is known as a long shot. (The dependent variable, which is not

TABLE 6.2    EXAMPLE OF A 2 X 2 FACTORIAL DESIGN TABLE			
		SMILING FACTOR	
		Smiling	Not Smiling
CAMERA DISTANCE FACTOR	Close-Up	<i>n</i> = 40	<i>n</i> = 40
	Long Shot	<i>n</i> = 40	<i>n</i> = 40

N = 160

shown in a table, could be something like how caring and compassionate subjects rate the person after seeing the photograph.) The rows and columns represent the factors; the cells represent one of the four distinct conditions, and this is described in the text of the article as: Smiling/Close-up; Smiling/Long shot; Not Smiling/Close-up; Not Smiling/Long shot. Inside the cells is the number of subjects in each condition, represented by the symbol *n*. The large N refers to the total number of subjects, the sum of all the cell *ns*. How to determine the number of subjects is covered in chapter 8. The number of conditions is formed by combinations of the levels of the independent variables and is found by multiplying the number of levels. In a 2 × 2 design, there are four conditions; in a 3 × 2 design, there are 6.

The cells of design tables can include other information, such as the marginal means of the groups once the study is conducted. Design tables can include as many levels of the two factors as needed simply by adding to the columns or rows, as shown in table 6.2. However, it can only accommodate a design with two factors because it is obviously two-dimensional; a three-dimensional table would be needed for three or more factors. It makes it easy to visualize how the number of subjects and stimuli increase as more levels are added.

TABLE 6.3    EXAMPLE OF A 2 X 3 FACTORIAL DESIGN TABLE			
		SMILING FACTOR	
		Smiling	Not Smiling
CAMERA DISTANCE FACTOR	Close-Up	<i>n</i> = 40	<i>n</i> = 40
	Medium Shot	<i>n</i> = 40	<i>n</i> = 40
	Long Shot	<i>n</i> = 40	<i>n</i> = 40

N = 240

## Choosing a Design

There is nothing about a factorial design that makes it inherently better or more theoretical than a single-factor design. However, if the topic being studied is likely to be influenced by two or more factors at a time, and if those factors are likely to interact, then a factorial design is a better choice. Factorial designs are more common than single-factor designs in journals, and the advantages of being more economical as well as being able to test interactions makes them more compelling. Bausell<sup>33</sup> points out that the more treatment groups there are, the more subjects are needed and the lower the probability of reaching significance when multiple statistical tests are performed. He says, “I would always counsel beginning with a two-group design.”<sup>34</sup> He then goes on to say that if the resources are available, more than two groups can add to theory by teasing out causal agents.

Researchers make decisions about which type of design to use based on evidence and what theory predicts about the concepts used as the dependent variable. Experiments that test two competing theories, or predictions that contradict each other, call for factorial designs. For example, in a study of how people learn about foreign nations from the media, two competing ideas about the causal mechanism of this type of influence were explored.<sup>35</sup> Logic can also dictate whether a study is a single-factor or factorial design. For example, a dissertation looked at a serious problem working journalists face in the current media environment: whether it is better to report a story quickly but with some inaccuracies, or to have it be accurate but lag behind all the other news organizations.<sup>36</sup> The speed vs. accuracy question lent itself to a single-factor design with two levels: fast with some inaccuracies, or slow but accurate. The reason for not splitting speed into two levels of speed (fast, slow) and accuracy (accurate, inaccurate) into separate levels is because the combinations of fast and accurate or slow and inaccurate are simply not realistic. Fast and accurate is obviously the ideal; if journalists could always produce this combination, there would be no problem. Slow and inaccurate is also an unlikely combination in the real world. If journalists are producing a story slowly, there is no excuse for inaccuracies; they have plenty of time to make sure information is right. Two of the four conditions that would result from a 2 x 2 factorial design simply did not make sense. Shadish, Cook, and Campbell call it a waste of resources “to test treatment combinations that are of no theoretical interest or unlikely to be implemented in policy.”<sup>37</sup> This illustrates the decision-making process; articles do not always explain why the researcher chose a single-factor or factorial design. However, the choice should always be made for good practical or theoretical reasons that can be explained to readers. In addition, the number of subjects rises with every additional factor and level, raising the study’s cost and time to conduct. Astronomical costs are a legitimate reason to limit factors. Practical things such as number of subjects, costs, study duration, and dropout rates, among others, are important to consider when choosing a design.<sup>38</sup> Because experiments are designed to



answer very specific questions raised by theory, narrowly focused studies can do a better job at developing theory than including “everything but the kitchen sink.” Common parlance is to say that including concepts X, Y, and Z were “beyond the scope of this study.”

### STUDY SPOTLIGHT 6.3

#### Example of a Single-Factor and Factorial Design in One Study

This study from organizational communication looked at how nonverbal expressions by spokespersons in times of crisis affect the public’s perceptions of the company. The authors did two experiments, one using a single-factor design and the other a factorial design.

In the first study, the single factor was voice pitch of the spokesperson, which the researchers manipulated to be either high or low. Subjects were assigned to only one condition—that is, they heard either the high-pitched voice or the low-pitched voice, not both—making this a between-subjects design.

In the second experiment, the researchers manipulate a different form of nonverbal communication than the verbal cues of the first study—visual cues—and they add a second factor, the stage of the crisis. This was a 2 x 2 factorial design with the first factor being the crisis stage and the two levels being during the crisis and after the crisis. The second factor was nonverbal visual cues, with the two levels being powerful or powerless. Read the study to see how power was manipulated in three nonverbal expressions. This resulted in four treatment conditions, diagrammed in this design table:

For Review-No Commercial Use(2023)

TABLE 6.4 ■ CRISIS STAGE FACTOR

Nonverbal Visual Cue Factor	Crisis Stage Factor		
		During	After
	Powerful	<i>n</i> = 29	<i>n</i> = 30
	Powerless	<i>n</i> = 30	<i>n</i> = 29

Note: The study does not say how many subjects were in each group, only that there were 118 total, so this is a guesstimate of the cell *ns*.

Subjects were randomly assigned to one of the four conditions, making this a between-subjects design. The literature review established that visual cues alone are known to have an impact on perceptions of the person’s competence, so this second experiment was not interested in the main effects of the two factors so much as the interaction of them. It found a significant interaction, showing that during a crisis, looking powerful leads to being perceived as competent, but that after the crisis, visual cues of power had no effect on perceptions. The study also tests perceptions of sincerity and other hypotheses but has been simplified here to illustrate the use of single-factor and factorial designs.

From: Claeys, An-Sofie, and Verolien Cauberghe. 2014. “Keeping Control: The Importance of Nonverbal Expressions of Power by Organizational Spokespersons in Time of Crisis.” *Journal of Communication* 64: 1160–1180.

## HOW SUBJECTS ARE USED IN DESIGNS

---

Another aspect of designing a true experiment is determining how many levels of a factor subjects will be exposed to. Described next, these also affect how many subjects are needed, how many stimuli must be created, and other concerns.

### Between-Subjects Designs

A **between-subjects design** is probably the easiest to conceptualize and one of the most commonly used. In this design, each subject is assigned to receive one and only one type of treatment, represented by one cell in the design table. Each subject is part of one group, not multiple groups. Morton and Williams call it being “in one state of the world.”<sup>39</sup> It is sometimes referred to as a “nested design” because subjects are nested within the levels of the condition.<sup>40</sup> The simplest between-subjects experiment is a single-factor design with two groups: a treatment and a control. A hypothetical example would be an education study with a treatment group that learned math using self-paced computer tutorials and a control group that learned from a teacher lecturing to a class, the usual way math is taught. In this case, a between-subjects design is necessary in order to avoid **carry-over effects**—that is, if subjects are likely to react to receiving the second condition because they have already received the first. Students can only be put in either the computer tutorial or lecture group because they would obviously learn the information, which would carry over to contaminate the outcome of being in the other group. Once they have learned the math, they cannot “unlearn” it. A completely different set of people is needed for each of the groups. Unmeasured variables that could possibly confound the results are controlled for with random assignment in a between-subjects design.

Between-subjects factors are also used when the condition is something the researcher cannot manipulate, like gender—a subject can only be in one gender. If the researcher in the hypothetical example thought that boys learned math differently from girls, for example, then gender might be another between-subjects factor. In one study, researchers measured adolescents’ prior drug use and used that as a factor, albeit one they could not manipulate.<sup>41</sup> Age, ethnicity, political ideology, and education, among others, are commonly used in experiments as between-subjects factors. These nonmanipulated factors are also called “measured factors.”

One drawback of the between-subjects design is the number of subjects necessary. If a power analysis shows the two groups in the math study need forty subjects in each cell, then eighty different subjects are needed. If a second type of treatment such as group learning is added, then another forty subjects are needed for a total of 120.

And this is just for a single-factor design. For a 2 x 2 factorial design that calls for forty subjects in each cell, that would mean a total of 160 subjects. A 2 x 3 design would require 200 subjects, and so on. The more complicated the design, the more subjects needed.

## Within-Subjects Designs

One way to get around the exponentially increasing number of subjects is to use a **within-subjects design**. If carry-over effects, such as doing better because of practice, are not an issue, this is the most economical type of design because the same subject can be part of all the treatment groups. Using the earlier example of a study<sup>42</sup> of news frames with two levels—objective and advocacy—there is no reason to think that reading an advocacy framed story would carry over to contaminate the response to reading an objective framed story or vice versa (as long as the stories are different), so a within-subjects design would allow the forty people to be included in both groups, whereas a between-subjects design would need eighty people, forty for each group. Morton and Williams call this being in “two states of the world simultaneously.”<sup>43</sup> It is sometimes called a crossed design because subjects are crossed with condition.<sup>44</sup> The same exponential increase is found when adding factors or levels. For example, in the hypothetical study of smiling and camera distance, a 2 x 2 design, each person could get all four types of photographs—one of a smiling close-up photo, one of a smiling long-distance photo, one of a not-smiling close-up photo, and one of a long-distance not-smiling photo. Because each person is represented in all four cells, a within-subjects design only needs forty people total rather than 120 for a between-subjects design. Essentially, each person “counts” as four subjects. It is sometimes called a “repeated measures” design. A system of randomly rotating the order the stimuli will be presented in should be employed to reduce potential carry-over effects; for example, a quarter of the subjects are randomly assigned to get smiling close-up photos first, a quarter to get smiling long-distance photos first, a quarter to get nonsmiling close-ups first, and a quarter to get nonsmiling long-distance photos first. This helps “wash out” any carry-over effects across the entire sample and will be covered in chapter 7.

The trade-off for within-subjects designs is that the researcher needs to create more stimuli. For example, in the study of objective and advocacy framed stories, the researcher could not use the exact same story written both ways. The researcher would need to write equivalent stories but have them be different enough so that subjects would not perceive they were reading the same thing twice. Where there is savings in subjects, there is more time spent creating stimuli.

The primary reason for not using a within-subjects design is to avoid carry-over effects—that is, if subjects are likely to react to the other conditions because of receiving the first one.

The easiest-to-understand example of this is if the “treatment” is really something that can cure a person. So if a factor is the type of headache medicine with three levels—aspirin, ibuprofen, and NSAIDs—then a researcher can only assign subjects with headaches to one cell or treatment because after taking the medicine their headache might be cured. Another way to think about it is if a subject cannot be returned to the way they were originally, then a within-subjects design is not appropriate.<sup>45</sup>

As a social science example, a study examined people’s belief in the ability of humans to have a free will to see how it would affect their charitable giving.<sup>46</sup> The free-will factor was manipulated by having subjects read an article from a science journal that argued against free will; the control group read an article on sustainable energy from the same science journal with no free-will messages. The dependent variable was how much subjects would give to charity, with those reading the article arguing against free will hypothesized to give less. This was a between-subjects design because once a participant read the article saying humans did not have free will, there was no way that could *not* affect their giving decisions. They could not be restored to their previous state where they did not know what the arguments were against humans having free will. They had been permanently changed. In other words, there is no going back from some treatments. The researchers could not have made this a within-subjects design without contaminating the treatments and confounding the results. Another example comes from a business study where the treatments were a company that voluntarily recalled a defective product versus a company that did not issue a recall. Subjects read about the companies and then rated them on integrity, trust, and benevolence.<sup>47</sup> Subjects could only be exposed to the voluntary recall or no recall condition, not both, as being exposed to the treatment in one would have an effect on the other.

Researchers also use between-subjects designs when they are concerned that subjects may “catch on” to the hypotheses. For example, after reading about recalls more than once in the business study, subjects may begin to suspect that researchers expect whether a company recalls a product to make a difference in how they view the company. We know from demand characteristics (see chapter 4) that subjects like to please researchers and may rate the recalling company better because they think that is what researchers want them to do. If it makes sense that subjects are assigned to only one kind of treatment, then a between-subjects design is called for. Another reason not to do a within-subjects design is if doing multiple tests takes a long time, subjects may become fatigued and not give the same amount of effort and attention to the later tests. However, if there is no reason to believe that subjects cannot receive more than one treatment without it affecting their responses, then the greater efficiency of within-subjects designs makes them preferable to between-subjects designs.

Another benefit of within-subjects designs is that each subject acts as his or her own control. That is, whatever individual differences the person has that may affect their response will be the same for both treatments. For example, in the advocacy vs. objective framing study,<sup>48</sup> having a college education might make a person think critically about all messages more than people who do not have a college degree. In that case, the college-educated subject is exposed to both advocacy and objective framed messages, which controls for education because each subject is acting as a control for him- or herself by getting both types of messages. In a between-subjects design, random assignment is used to help control for individual differences by helping ensure that experimental groups are equivalent in terms of things like education, age, gender, race, and other characteristics that might make a difference—for example, political ideology or even psychological characteristics such as fundamentalism or authoritarian personalities. But equivalence of groups using random assignment is never as good as equivalence achieved by having exactly the same subject in each group. This will be covered in more detail in chapter 7.

## Mixed Factorial Designs

Some studies use a combination of between- and within-subjects designs in the same experiment. Within these designs, one factor is a between-subjects factor and the other is a within-subjects factor. When there are more than one factor, and one is a between-subjects factor and the other is a within-subjects factor, then a **mixed factorial design** is appropriate. In these designs, one independent variable is given as a between-subjects factor so that subjects get only one level of that treatment, and the other is given as a within-subjects factor so that subjects get all levels of that treatment. For example, in one study that used a mixed design,<sup>49</sup> subjects read four ethical dilemmas that journalists might face and then made a decision about what they would do. The within-subjects factor was the type of decision they had to make, with the two levels being whether they should run a photograph or not and whether they should run a story with anonymous sources or not. All subjects got two dilemmas where they made decisions about running photographs and two dilemmas where they made decisions about using anonymous sources. The between-subjects factor was whether subjects got photographs with the dilemmas or not. For this factor, half of the subjects saw photographs with all four dilemmas, and the other half got all four dilemmas without any photographs. This was done because making a decision about running photographs was not expected to have any effect on subjects' decision to use anonymous sources or not, and vice versa; however, there was concern that seeing photographs for some dilemmas might color the way subjects reacted to other dilemmas where they did not see photographs. It made sense for subjects to receive only one of the levels of the photograph factor; they either received photographs or did not receive photographs. In this between-subjects factor, each subject

got one kind of treatment, not both. It also made sense for them to receive both kinds of decisions, using photographs or anonymous sources. In this within-subjects factor, each subject got both treatments.

At first, it may be hard to remember which type of design has subjects in only one group and which has them in both. Try this mnemonic phrase: “Choose *between* the two; you can only have one.” Thus, a between-subjects design has subjects in only one group.

The paper should report which type of subjects’ design was used—between, within, or mixed—along with the factorial notation. Here are examples, with the type of subject design underlined:

- “We employed a 2 (argument quality: strong/weak) x 2 (personal relevance: high/low) x 2 (evidence: narrative/statistical) mixed factorial design.”<sup>50</sup>
- “The second study used a 2 (crisis stage: during the crisis vs. postcrisis) x 2 (visual nonverbal cues: powerless vs. powerful) between-subjects factorial design.”<sup>51</sup>

For more examples of how to report experimental designs in the paper, see How To Do It box 6.2, Reporting Factorial Design.

For Review-No Commercial Use(2023)

## Incomplete Factorials

All the designs described previously are what are called **fully crossed factorial designs**, sometimes shortened to just **full factorials** or **complete factorials**. In these, all combinations of the levels and factors are assigned subjects. But some designs leave one or more of the cells in the design table empty; these are called **incomplete factorial designs** or **fractional factorial** designs. As we discussed earlier, some of the combinations might not make sense. Shadish, Cook, and Campbell give an example of a research design that did not assign subjects to eight cells that were either too politically unpopular or expensive to be able to implement in a government policy on poverty.<sup>52</sup> When there is no practical or theoretical reason to study some cell combinations, it is more economical to leave those cells empty and not assign any subjects to them; then it is reported as an incomplete factorial design. Another reason to use an incomplete factorial is to allow for a control group that receives no treatment. If a control group does not fit into any of the cell combinations, it is considered to be its own cell. It would look like the design in table 6.5. The hypothetical study in the figure has three levels of the “visual type” factor, where subjects see either a still photograph, watch a video shown one time the way it is on TV, or watch a video shown three times, the way one could watch it on the Internet. The second factor represents the camera distance from the people in the images, either a close-up, medium, or a long shot. The control group got no pictures or video, only the story that all the other subjects got with

**TABLE 6.5**    **EXAMPLE OF A 3 X 3 INCOMPLETE FACTORIAL DESIGN WITH A CONTROL GROUP**

		VISUAL TYPE FACTOR		
		Still Photo	TV Video (1x)	Internet Video (3x)
CAMERA DISTANCE FACTOR	Close-up	n = 45	n = 45	n = 45
	Medium shot	n = 45	n = 45	n = 45
	Long shot	n = 45	n = 45	n = 45

Control Group (no image treatment/stories only)

n = 15

their visuals. There was no way to “fit” this into the design table because both factors in the table were a feature of photographs or video. Thus, the control group gets its own cell, outside the table. This kind of experiment is more complicated because the control group can have fewer subjects than the treatment groups. At least a 3:1 ratio is recommended,<sup>53</sup> meaning that up to three times as many subjects can be in the treatment groups as the control.

## CONTROL GROUPS

One of the advantages of a factorial design, whether it is complete or incomplete, is for control in the form of a comparison to a group that does not receive a treatment.<sup>54</sup> In complete factorials, one of the levels of the factors can represent the control condition<sup>55</sup>—for example, a study<sup>56</sup> that was a 4 x 3 mixed factorial with one factor called Social Issue, represented by three different stories, one about a flood victim, one about teenage girls beating another girl, and one about a prostitution sting. The second factor was visual type, the same as in the hypothetical example—still photo, TV video, Internet video—but with a control group: No Image. Control group subjects read the same stories about the three different social issues but saw no pictures or video. That design table looked like table 6.6.

As we see from Campbell and Stanley’s<sup>57</sup> typology of experimental designs covered in chapter 4, a control group is essential to all three of the “true experiments.” The pretest–posttest control group design and the posttest-only control group design both have one control group, while the Solomon four-group design has two. A control group is crucial in ensuring that different groups are reacting under controlled conditions.<sup>58</sup>

TABLE 6.6    **EXAMPLE OF A 4 X 3 FACTORIAL DESIGN WITH A CONTROL GROUP**

		VISUAL TYPE FACTOR			
		Still Photo	TV Video (1x)	Internet Video (3x)	No Image (Control Group)
SOCIAL ISSUE FACTOR	Beating	n = 33	n = 45	n = 38	n = 29
	Flood	n = 33	n = 45	n = 38	n = 29
	Sting	n = 33	n = 45	n = 38	n = 29





Another way to look at a control group is as a baseline measure of something that has not yet had a treatment or intervention,<sup>59</sup> or as a neutral situation.<sup>60</sup> If the experimental group receives the treatment and changes, and if the control group does not receive the treatment and changes, then we know the treatment did not cause the change. But if the experimental group receives the treatment and changes, while the control group that does not receive the treatment does not change, then we know the treatment did cause the change.<sup>i</sup>

After random assignment, having a control group is the most important thing one can do when conducting a true experiment. It is the way experimenters compare what happens to people who get the treatment against people who do not. The control group is another way of ensuring that any effects observed are due to the treatment variable and not something else. As pointed out in chapter 1, medical researchers typically give their control group subjects a placebo such as a sugar pill, injection of saline, or something else that leads people to believe they got some “treatment” even if it was inert or not expected to have any beneficial effects. Because of the power of suggestion, people need to think they have gotten some treatment or had something done to them in order to accurately represent a control group. As Campbell and Stanley<sup>61</sup> point out, thinking of a control group as “the comparison of *X* with *no X* is an oversimplification. The comparison is actually with the specific activities of the control group which have filled the time period corresponding to that in which the experimental group receives the *X*.”<sup>62</sup> An adequate control group is not a group of people who are measured on some outcome variable without having received something that they perceive to be a treatment. For social scientists, there is no “sugar pill.” Next, this chapter discusses some ways that these researchers create social science placebos.

## The No-Treatment Control Group

The classic example of a control group is one that gets the absence of the manipulation in the treatment group. For example, researchers in one study used stories that explicitly called an issue “important” for one treatment group and stories that implicitly labeled an issue as important by presenting its consequences for the other treatment group. The control group saw the exact same stories without the evaluation of the issue as being important or any discussion about consequences.<sup>63</sup> In the insecurity study example discussed earlier, the treatment group wrote essays about their financial futures, creating feelings of high insecurity. The control group wrote essays about music, which was not expected to

<sup>i</sup>I would like to thank an anonymous reviewer for suggesting this clear and simple way of describing control groups.

create any feelings of insecurity.<sup>64</sup> Key to the idea of control is that subjects in the control or “no treatment” group get a manipulation that is as similar to the treatment group as possible without affecting the outcome.

## Creativity in Control Groups

Sometimes, it takes a little creative thinking to determine what subjects should do instead of doing nothing. For example, researchers studying cooking practices used three types of marketing messages as their treatment: messages about health benefits, messages about time and money savings, and messages about both.<sup>65</sup> No message was the obvious control group, but then what would they have subjects do instead? The researchers solved that problem by using a discussion group as the control condition, having subjects talk with others about common cooking practices instead of being exposed to a marketing message. In a journalism study that compared advocacy to objective stories about crime, the control group read stories about topics other than crime so they would not experience the same emotions, such as fear, that crime stories engender.<sup>66</sup> They still read stories, just not crime stories. In both of these studies, instead of having the control group subjects do nothing, they did something—participate in a group discussion instead of seeing a marketing message, or read a newspaper story on a topic completely unrelated to the topic of the treatment groups' stories. What Bausell calls “jazzing up the control group to receive interesting but irrelevant activities”<sup>67</sup> means that giving control subjects something that appears to be the same as the treatment subjects is much better than having them come in, fill out the questionnaire, and then leave, without getting anything as a treatment. Control group subjects must perform some task that makes them think they have received a legitimate treatment. They need to perceive that something happened.

## The Status Quo Control Group

In other cases, researchers will define the control group as what is most normal, common, or typical.<sup>68</sup> In these experiments, a different group that represents the status quo but is not identified as a control group may be used. For example, in one experiment of how to best teach English to Costa Rican third-graders, two treatment groups used different computer-assisted learning software while a control group got English instruction by a teacher, the way the schools normally taught it.<sup>69</sup> The teacher-instructed group served as the control group. Journalism studies frequently use stories written in the detached, two-sided manner that is common in mainstream news organizations as the control condition because it represents what audiences normally read.<sup>70</sup>

## No Control Group

Not all experiments have control groups, and in some cases one may not be necessary.<sup>71</sup> In some studies, it is extremely difficult to provide a control treatment that is similar to the manipulation. For example, in the study of HIV messages, what would be the control for the story frame factor?<sup>72</sup> There is no such thing as a story that does not have some sort of frame. Or in the study of voice pitch, where the two levels were raised and lowered voice pitch<sup>73</sup>—there is no such thing as no voice pitch. In many studies that compare two or more treatments, what is defined as the control group is ambiguous.<sup>74</sup> In these studies, a control group may not be necessary because they are comparing the treatment groups to each other,<sup>75</sup> not against some group that gets the absence of voice pitch or a story with no frame.

Whether a study has a classic control group or some other kind, there should always be careful control in the management of variables.<sup>76</sup> Refer to the threats to validity outlined in chapter 5 when designing an experiment and put every element under a microscope.<sup>77</sup> Groups should differ only with respect to the intervention. Even something as simple as running one group in the morning and the other in the afternoon, or in dissimilar rooms, or with different experimenters, can confound the outcome.<sup>78</sup>

## For Review-No Commercial Use(2023)

There are, of course, more types of factorial designs than the ones outlined here, notably split plot designs, fractional replications, and randomized blocks, among others—what Campbell and Stanley call “some of the traumatizing mysteries of factorial design.”<sup>79</sup> As this is an introductory text, it will not delve into the mysteries of each but instead refer readers to more advanced texts.

The next chapter will take on the single most important thing that must be done to qualify as a true experiment—random assignment.

### Common Mistakes

- Designing a study that is too complicated
- Not using a control group when possible
- Not giving the control group something to do

## Test Your Knowledge

1. An experiment that has two independent variables with two or more levels of each is called a \_\_\_\_\_.
  - a. Single-factor design
  - b. One-way design
  - c. Factorial design
  - d. Three-way design
2. What factorial notation would be used for this hypothetical experiment?: A researcher manipulates the tone of an ad campaign (positive/negative/neutral) and the medium that carries it (TV/print/Internet).
  - a.  $2 \times 2$
  - b.  $2 \times 3$
  - c.  $3 \times 3$
  - d.  $2 \times 3 \times 3$
3. A factorial design can have which of the following outcomes?
  - a. Significant main effects for all the factors but no interaction effect
  - b. Significant main effects for one or some of the factors but not all of them, with either a significant or nonsignificant interaction
  - c. A significant interaction but no main effects
  - d. All of these
4. If subjects are likely to do better on something because of practice, then it is best to use a(n) \_\_\_\_\_.
  - a. Between-subjects design
  - b. Within-subjects design
  - c. Mixed design
  - d. Incomplete design
5. Identify the control group in a study of the best way to remind drivers to pay a traffic ticket:
  - a. Reminders sent via text message
  - b. Reminders sent by e-mail
  - c. Reminders sent by mail, the way they are normally sent
  - d. Reminders sent by phone call
6. In a study on PSAs about HIV/AIDS, researchers examined the quality of the argument in the ad with the levels being a strong or weak argument, how personally relevant the subject of HIV/AIDS was to each subject, and two different types of ad formats, one that used a narrative storyline and one that relied on statistical data. This was a \_\_\_\_ design.

- a. Single-factor
  - b. Factorial
  - c. Between-subjects
  - d. Within-subjects
7. The ability for the different levels in one factor alone to cause a change in the outcome variables is called a(n) \_\_\_\_.
- a. Confound
  - b. Plausible alternative explanation
  - c. Main effect
  - d. Interaction effect
8. When the outcome of one factor depends on another, this is known as a(n) \_\_\_\_.
- a. Confound
  - b. Plausible alternative explanation
  - c. Main effect
  - d. Interaction effect
9. In a study of cooking practices where the three conditions were messages about health benefits, messages about time and money savings, messages about both, and a discussion group, which represented the control group?
- a. Messages about health benefits
  - b. Messages about time savings
  - c. Messages about money savings
  - d. A discussion group
10. In an education study that compared computer-assisted learning software to instruction by a teacher, the way the schools normally taught it, which kind of control group was used?
- a. No control group
  - b. Status quo control group
  - c. Comparison of two treatments
  - d. Internal control group

Answers:

- |      |      |      |       |
|------|------|------|-------|
| 1. c | 4. a | 7. c | 9. d  |
| 2. c | 5. c | 8. d | 10. b |
| 3. d | 6. b |      |       |

## Application Exercises

1. Take one of the ideas for an experiment that you came up with in chapter 1 and wrote a literature review and hypotheses for in chapter 3, and draw a design table to help you visualize the factors, levels, and the combinations of each. Translate that into factorial notation. Include the type of design, your factors, and levels. See the How To Do It box.
2. Write one page that explains your choice of between- or within-subjects designs. Defend your decision (e.g., why there will be no carry-over effects). Imagine a reviewer has challenged your choice, suggesting that the opposite would be better. What would you say?
3. Write one page on your type of control group and what the subjects will do instead of receiving the treatment. Explain your choices and defend your decisions.

## Suggested Readings

Chapter 6, “Designing Experimental Studies to Achieve Statistical Significance,” in R. Barker Bausell (1994). *Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist*. Thousand Oaks, CA: Sage Publications.

### For Review-No Commercial Use(2023)

Chapter 10, “Complex Experimental Designs,” in Paul C. Cozby (2009). *Methods in Behavioral Research*, 10th Edition. Boston: McGraw-Hill.

## Notes

1. Diana Mutz, *Population-Based Survey Experiments* (Princeton, NJ: Princeton University Press, 2011), 125.
2. Graeme D. Ruxton and Nick Colegrave, *Experimental Design for the Life Sciences*, 3rd ed. (Oxford, UK: Oxford University Press, 2011).
3. William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Belmont, CA: Wadsworth Cengage Learning, 2002), 263.
4. Sean Aday, “The Framesetting Effects of News: An Experimental Test of Advocacy Versus Objectivist Frames,” *Journalism and Mass Communication Quarterly* 83, no. 4 (Winter 2006): 767–784.
5. Tim Kasser and Kennon M. Sheldon, “Of Wealth and Death: Materialism, Mortality Salience, and Consumption Behavior,” *Psychological Science* 11, no. 4 (July 2000): 348–351.
6. Christopher D. Johnston and Julie Wronski, “Personality Dispositions and Political Preferences Across Hard and Easy Issues,” *Political Psychology* 36, no. 1 (2015): 35–53.
7. T. Rogers et al., “Artful Paltering: The Risks and Rewards of Using Truthful Statements to Mislead Others,” *Journal of Personality and Social Psychology* 112 (3), 456–473.
8. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
9. Ronald A. Fisher, *The Design of Experiments* (Edinburgh: Oliver and Boyd, 1937).
10. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. (Chicago: Rand McNally, 1963), 3.

11. Ibid.
12. Jueman Zhang et al., "Persuasiveness of HIV/AIDS Public Service Announcements as a Function of Argument Quality, Personal Relevance, and Evidence Form," *Social Behavior and Personality: An International Journal* 42, no. 10 (2014): 1603–1612.
13. Ruxton and Colegrave, *Experimental Design for the Life Sciences*, 3rd ed.
14. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*, 29.
15. Ruxton and Colegrave, *Experimental Design for the Life Sciences*, 3rd ed.
16. Zhang et al., "Persuasiveness of HIV/AIDS Public Service Announcements."
17. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*, 29.
18. Ibid.
19. Xiaoli Nan et al., "Message Framing, Perceived Susceptibility, and Intentions to Vaccinate Children Against HPV Among African-American Parents," *Health Communication* 31, no. 7 (2016): 798–805.
20. Ibid.
21. Gabriella Esposito et al., "Judging to Prevent the Purchase of Incompatible Digital Products Online: An Experimental Study," *PLoS ONE* 12, no. 3 (2017): 1–15.
22. Nan et al., "Message Framing, Perceived Susceptibility."
23. Ibid., 800.
24. Ruxton and Colegrave, *Experimental Design for the Life Sciences*, 3rd ed.
25. Debby Vos, "How Ordinary MPs Can Make It Into the News: A Factorial Survey Experiment with Political Journalists to Explain the Newsworthiness of MPs," *Mass Communication and Society* 19, no. 6 (2016): 738–757.
26. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 265.
27. Ibid.
28. R. Barker Bausell, *Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist* (Thousand Oaks, CA: Sage, 1994).
29. Murray R. Selwyn, *Principles of Experimental Design for the Life Sciences* (Boca Raton, FL: CRC Press, 1996), 55.
30. Zhang et al., "Persuasiveness of HIV/AIDS Public Service Announcements," 1606.
31. An-Sofie Claeys and Verolien Cauberghe, "Keeping Control: The Importance of Nonverbal Expressions of Power by Organizational Spokespersons in Time of Crisis," *Journal of Communication* 64 (2014): 1160–1180.
32. Ibid., 1162.
33. Bausell, *Conducting Meaningful Experiments*.
34. Ibid., 97.
35. Paul R. Brewer, Joseph Graf, and Lars Willnat, "Priming or Framing," *Gazette: International Journal for Communication Studies* 65, no. 6 (2003): 493–508.
36. Angela Min-Chia Lee, "How Fast Is Too Fast? Examining the Impact of Speed-Driven Journalism on News Production and Audience Reception" (unpublished dissertation, University of Texas, 2014); Angela M. Lee, "The Faster the Better? Examining the Effect of Live-Blogging on Audience Reception," *Journal of Applied Journalism and Media Studies* 8, no. 1 (in press).
37. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 265.
38. Selwyn, *Principles of Experimental Design*, 53.
39. Rebecca B. Morton and Kenneth C. Williams, *Experimental Political Science and the Study of Causality: From Nature to the Lab* (New York: Cambridge University Press, 2011), 80.
40. Charles M. Judd, Jacob Westfall, and David A. Kenny, "Experiments with More Than One Random Factor: Designs, Analytic Models, and Statistical Power," *Annual Review of Psychology* 68, no. 1 (2017): 601–625.
41. H. Cho and F. J. Boster, "Effects of Gain Versus Loss Frame Antidrug Ads on Adolescents," *Journal of Communication* 58, no. 3 (2008): 428–446.
42. Aday, "The Framesetting Effects of News."
43. Morton and Williams, *Experimental Political Science*, 86.
44. Judd, Westfall, and Kenny, "Experiments with More Than One Random Factor."
45. Ruxton and Colegrave, *Experimental Design for the Life Sciences*, 3rd ed.
46. Job Harms et al., "Free to Help? An Experiment on Free Will Belief and Altruism," *PLoS ONE* 12, no. 3 (2017): 1–15.
47. Luiza Venzke Bortoli and Valeria Freundt, "Effects of Voluntary Product Recall on Consumer's Trust," *Brazilian Business Review* 14, no. 2 (2017): 204–224.
48. Aday, "The Framesetting Effects of News."
49. Renita Coleman, "The Effect of Visuals on Ethical Reasoning: What's a Photograph Worth to Journalists Making Moral Decisions?" *Journalism and Mass*

- Communication Quarterly* 83, no. 4 (Winter 2006): 835–850.
50. Zhang et al., “Persuasiveness of HIV/AIDS Public Service Announcements,” 1606.
  51. Claeys and Cauberghe, “Keeping Control,” 1167.
  52. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 265.
  53. Bausell, *Conducting Meaningful Experiments*.
  54. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
  55. Ibid.
  56. Aimee Meader et al., “Ethics in the Digital Age: A Comparison of the Effects of Moving Images and Photographs on Moral Judgment,” *Journal of Media Ethics* 30, no. 4 (2015): 234–251.
  57. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  58. Gerry Stoker, “Exploring the Promise of Experimentation in Political Science: Micro-Foundational Insights and Policy Relevance,” *Political Studies* 58 (2010): 300–319.
  59. Morton and Williams, *Experimental Political Science*.
  60. Cengiz Erisen, Elif Erisen, and Binnur Ozkececi-Taner, “Research Methods in Political Psychology,” *Psychological Studies* 13, no. 1 (2013): 13–33.
  61. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*.
  62. Ibid., 13.
  63. Kristin Bulkow, Juliane Urban, and Wolfgang Schweiger, “The Duality of Agenda-Setting: The Role of Information Processing,” *International Journal of Public Opinion Research* 25, no. 1 (2013): 43–63.
  64. Kasser and Sheldon, “Of Wealth and Death.”
  65. Theresa Beltramo et al., “The Effect of Marketing Messages and Payment Over Time on Willingness to Pay for Fuel-Efficient Cookstoves,” *Journal of Economic Behavior & Organization* 118 (2015): 333–345.
  66. Aday, “The Framesetting Effects of News.”
  67. Bausell, *Conducting Meaningful Experiments*, 74.
  68. Ibid.
  69. Horacio Alvarez-Marinelli et al., “Computer Assisted English Language Learning in Costa Rican Elementary Schools: An Experimental Study,” *Computer Assisted Language Learning* 29, no. 1 (2016): 103–126.
  70. Aday, “The Framesetting Effects of News.”
  71. Ruxton and Colegrave, *Experimental Design for the Life Sciences*, 3rd ed.
  72. Nan et al., “Message Framing, Perceived Susceptibility.”
  73. Claeys and Cauberghe, “Keeping Control.”
  74. Morton and Williams, *Experimental Political Science*.
  75. Ruxton and Colegrave, *Experimental Design for the Life Sciences*, 3rd ed.
  76. Stoker, 304.
  77. Bausell, *Conducting Meaningful Experiments*.
  78. Ibid.
  79. Campbell and Stanley, *Experimental and Quasi-Experimental Designs*, 27.





# RANDOM ASSIGNMENT

*Just as representativeness can be secured by the method of chance . . . so equivalence may be secured by chance.<sup>1</sup>*

—W. A. McCall

## LEARNING OBJECTIVES

For Review-No Commercial Use(2023)

- Understand what random assignment does and how it works.
- Produce a valid randomization process for an experiment and describe it.
- Critique simple random assignment, blocking, matched pairs, and stratified random assignment.
- Explain the importance of counterbalancing.
- Describe a Latin square design.

Just as the mantra in real estate is “location, location, location,” the motto in experimental design is “random assignment, random assignment, random assignment.” This book has discussed random assignment all throughout. It bears repeating that random assignment is the *single* most important thing a researcher can do in an experiment. Everything else pales in comparison to having done this correctly.<sup>2</sup> Random assignment is what distinguishes a true experiment from a quasi, natural, or pre-experimental design. In chapter 1, experiments were referred to as the gold standard. Without successful random

assignment, however, they can quickly become “the bronze standard.”<sup>3</sup> This chapter will review some of the advantages of random assignment, discuss the details of how to do it, and explore related issues of counterbalancing.

## THE PURPOSE OF RANDOM ASSIGNMENT

---

People vary. That is, they are different. Were this not so, there would be no reason to study them. Everyone would be the same, reacting the same way to different teaching techniques, advertisements, health interventions, and political messages. There would be no need to conduct experiments. The fact that people vary provides social scientists with a reason for doing research, and also with their biggest challenge. When people vary on things that are not of interest to the experiment, it is called **random variation** or **noise**.<sup>4</sup> These things that are not the focus of a study can still be responsible for some of the changes in the outcome of experimental treatments. That is, they can confound the results. One major focus of experimentalists is to control for confounds, removing or reducing the noise from random variation. That way, the effects the study is concerned with can be seen more clearly. Randomization is arguably the greatest weapon a scientist has because it helps ensure that subjects in different treatment and control groups are virtually the same on variables that create noise.

Random assignment is a technique for placing subjects into the different treatment and control groups in a true experiment for the purpose of ensuring that subjects in each group will have similar characteristics—that is, that they will be **equivalent**. By equivalent, we mean equal on average, or probabilistically equal, not identical. The purpose of equivalence is to “level the playing field,” helping ensure that the only systematic differences between the two groups are the treatments they receive in the experiment. It helps ensure that subjects in one group are not better on the outcome variable to begin with, and that one group is not filled with subjects who are more likely to change regardless of treatment.<sup>5</sup> This allows researchers to have more confidence that any changes observed are because of the treatment the subjects received and not because of inherent differences in the subjects themselves. Groups that are systematically different in one or more ways can invalidate an experiment. For example, if one group contained only men and the other only women, it would be confounded; there would be no way to tell if the results were due to the treatment or to gender. One way random assignment helps achieve equivalence is by avoiding **selection bias**, which occurs when people self-select the groups to be in, or researchers select subjects for some subjective reason or to improve the chances of supporting a hypothesis, even if done subconsciously.<sup>6</sup>

For example, an education study would be invalid if the treatment group gets mostly subjects who are better at math to begin with and the control group gets subjects who are poor at math. If the treatment is the way math is taught, then it might appear as if the treatment is working, but it really could be the fact that those who were taught the new way were better at math to begin with. On the other hand, if students who are poor at math are assigned to the treatment group by educators who want to be sure that students who need math help the most get it, then the treatment group is stacked with poor-math students and the control group with good-math students. The new teaching strategy could actually work, but might look like it was not if it only raised the treatment group up to the level of the controls; in other words, there is no significant difference between the groups after treatment.

Another example is that some people are simply more prone to change than others. If the treatment group got more subjects who changed more easily than the control group, then a study designed to influence people's positions on public policy could show spurious results—the treatment did not really change people's minds; they were more likely to change their minds to begin with. Giving the same treatment to harder-to-change people might have no effect at all.

## For Review-No Commercial Use(2023)

### Avoiding Confounds

The beauty of random assignment is that the researcher does not have to predetermine every characteristic of every subject that could possibly confound the study. It might be easy to anticipate that prior math knowledge would influence math learning, for example, and pretest subjects on math and assign equivalent numbers of high-math knowledge and low-math knowledge subjects to each group. But it might be more difficult for a researcher to anticipate a penchant for changing one's mind as a confounding variable. With random assignment, no pretests for math ability or mind changing are necessary because equivalent numbers of easy-changers and hard-changers are in each group; or equivalent numbers of good-at-math and poor-at-math students are in each group. Because humans, including researchers, are notoriously bad at anticipating every little thing that might affect something else, and because some things are unknowable, random assignment is of tremendous benefit. Also, recall the drawbacks of pretesting from chapter 4; random assignment can eliminate the need for pretests and their accompanying threats to validity.

Random assignment, while not perfect, is the best way we currently know of to ensure that systematic variation among subjects does not confound the results of an experiment. It helps eliminate spurious variables, including those a researcher might not have thought of.

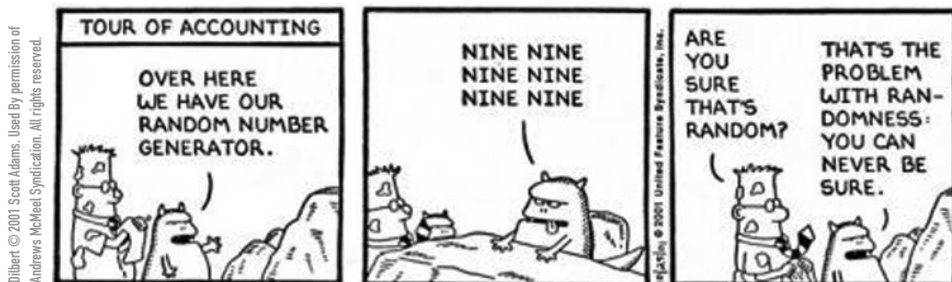
When it is impossible for the same people to simultaneously have the treatment and not have it, covered in chapter 6, researchers use random assignment to try to make sure that individual differences are assigned equivalently to each group.

## What Is Random?

Random assignment works because of chance. To assign subjects randomly, everyone in the study must have an equal chance of being in the control or treatment groups. Chapter 2 recounted the story of how random assignment was discovered. When Charles Sanders Peirce and Joseph Jastrow conducted a study of how accurately people could judge the weight of something just by feeling and looking at it,<sup>7</sup> they started by presenting the heavy weights first. Then they presented the heavy weights last. They also alternated the heavy and light weights. Finally, they shuffled a deck of cards and assigned the weights at random based on the cards. They got vastly different results when the weights were presented in any of the systematic patterns than when they were presented in a random order. Having subjects who could not guess the weight based on a pattern produced more valid results.

Random does *not* mean haphazard, and researchers must be careful to use appropriate random methods. What is not random is anything that has some kind of pattern, purpose, or system to it. In my first experiment, I apparently did not understand exactly what random meant, but I was concerned with having equal numbers of subjects in each group. I assigned the first person to come into the lab to the first group, the second person to the second group, then the third person back to the first group, etc. Basically, they were assigned like this: 1, 2, 1, 2, 1, 2, 1, 2 . . . . When I told my professor, her eyes got wide and she said, “That’s NOT random.” I ended up throwing away all that data and starting over again.

As Gueron says, “It does not help to be a little bit random.”<sup>9</sup>



## MORE ABOUT . . . BOX 7.1

### Random Assignment Threats to Internal Validity

When subjects and/or the experimenters that assign subjects to condition are not “blind” to the group they are being assigned to, random assignment itself, and violations of it, can result in threats to internal validity. The four types are:<sup>13</sup>

- *Diffusion of treatment*—This occurs when subjects in different groups communicate with each other and learn information not intended for them. For example, if subjects in the treatment and control group talk to each other, they may learn material intended only for the treatment group; the outcomes may not be different and will not truly reflect the treatment benefits.
- *Compensatory equalization*—This occurs when experimenters or those providing the treatment attempt to give some of the advantages that the treatment group has to those in the control group.
- *Compensatory rivalry*—This refers to when members of the control group try to gain some of the benefits of the treatment group.
- *Resentful demoralization*—This refers to the control group subjects underperforming because they resent being denied the treatment.

For Review-No Commercial Use(2023)

In some studies, subjects in the treatment group feel demoralized or stigmatized—for example, as being in the class for “dummies.”<sup>14</sup> Thus, it is important that subjects be blind to which group they are in whenever possible.<sup>15</sup>

Researchers should be careful not to portray one intervention group as better or newer than another in recruitment materials.<sup>16</sup> If it is not possible to keep subjects from knowing which group they are in, researchers should measure subjects’ preferences for assignment to a particular group and statistically control for it.<sup>17</sup>

Another threat to internal validity occurs when there is an imbalance of subjects who have a greater preference for their assignment in one group than in the other.<sup>18</sup> For example, if 60% of subjects in both groups are pleased with the group they are in and 40% are not, then preference is equivalent across all conditions. However, if one group has 60% pleased and the other has only 40% pleased, the outcomes could be confounded.<sup>19</sup> As long as both groups have the same percentage of pleased and displeased subjects, no threat occurs.<sup>20</sup> For example, in a business study, 89% of those asked to be in the treatment group agreed to participate, whereas only 45% of those asked to be in the control group agreed.<sup>21</sup>

True randomization is a process that is either done correctly or not.<sup>10</sup> In order for random assignment to work, the researcher cannot choose which group a subject is in for any reason, nor can a subject choose his or her own group. The subjects must be **blind**, or unaware, as to whether they are receiving the treatment or not. It is important that subjects

remain in the dark so they do not try to give researchers what they think they want,<sup>11</sup> or so their disappointment at not getting the treatment does not bias the results.<sup>12</sup> It is also a good idea to have the researcher be unaware of which treatment or control groups subjects are in, which is then termed a *double-blind* study. For more about threats to internal validity when subjects and/or researchers know which group they are assigned to, see More About box 7.1.

## OPERATIONALIZING RANDOM ASSIGNMENT

---

There are many ways to “do” random assignment. In 1883, Peirce and Jastrow used a special deck of cards to determine random assignment, and that is still a valid method today. Other methods include rolling dice, flipping a coin, drawing numbers out of a hat, or using a book of random numbers. For details on how to do these manual methods, see How To Do It box 7.2.

### Computerized Randomization

Today, it is more likely that researchers will use random number generators found online or in spreadsheet or statistical software. For example, free online randomizers allow a researcher to specify the number of subjects and number of groups, and quickly return a list showing which subjects go to which groups, as shown in figures 7.1 and 7.2 (see pp. 180–181). Simply search the Internet for “random assignment generator” to find these.

To use QuickCalcs by GraphPad, in the first box labeled “Assign,” put in the total number of subjects. Put the number of groups in the second box. Leave “Repeat” at 1. Click “Do it!”

It will return a list of subjects, numbered 1 to your final number, with the group labeled A, B, C, etc., beside the subject number.

This applet is specifically designed by a group of academic researchers for factorial experiments. In the first box, “Number of Participants,  $N$ ” put the total number of subjects for the entire study. In the second box, “Number of Conditions,  $C$ ,” put the number of groups. Click “Compute.”

It will return a list of groups to assign participants to, in order; it lacks the subject number of the QuickCalcs randomizer, but it is not hard to see that the first subject is assigned to the first group number, the second subject to the group number listed next, etc.

## HOW TO DO IT 7.2

### Randomizing Subjects

*Coin flipping* works when there are two groups. If the coin lands heads up, assign the subject to the treatment group; if it lands tails up, assign the subject to the control group. Or vice versa.

A *lottery* works with any number of groups. On slips of paper, write a number for each of the groups (1, 2, 3, 4, for example). Make as many slips of paper as subjects you intend to have, with equal amounts of group numbers. For example, if you need 160 subjects, with forty in each group, make forty slips of paper with the number 1, forty with number 2, etc.

*Decks of cards and dice* will need to use cards and die with only the number of groups on them. For example, get rid of all cards that are not 1, 2, 3, or 4 if you have four groups. You might need multiple decks. Shuffle the cards and then draw them one at a time each time a subject arrives. The subject is assigned to the group represented on the card drawn. For dice, use only dice that have the same number on them as your groups. Or roll again if a number comes up that is greater than your number of groups. Once the maximum number of subjects in a group is reached, ignore that number when it comes up on a card or die.

*Book of random numbers.* These are obsolete today, replaced by online random number generators. They consisted of page after page of numbers, randomly ordered. Believe it or not, a researcher decided where to start by closing his or her eyes and pointing to a starting place on the table. Then, the researcher would assign subjects to the group that the numbers correspond to, skipping over numbers that are outside the range. For example, if the numbers are: 3, 2, 5, 3, 7, 4, 1, etc., and the study has three groups, subjects would be assigned to groups 3, 2, 3, 1, etc., skipping over 5 and 7 because there are not five or seven groups. Once the maximum number of subjects in a group has been reached, that group's number is skipped over as well. Stop assigning subjects to a group after a group reaches the maximum, but continue on until all the other groups have been filled. My favorite book of random numbers was *A Million Random Digits With 100,000 Normal Deviates* (RAND). Your library might still have it if you are curious or a history buff.

*Excel.* For a tutorial on how to draw random numbers in Excel, see <https://exceljet.net/formula/randomly-assign-data-to-groups>.

*SAS and SPSS.* Shadish, Cook, and Campbell<sup>22</sup> give directions on random number generation in these popular statistical software packages on pages 311–313.

For other statistical software, such as *Stata* and *R*, consult the tutorials.

For pencil-and-paper studies, an online randomizer can be used to arrange the printed questionnaires in the specified random order before going to the lab or site where subjects will take the study. Questionnaires can then be handed out from the top of the stack without needing to refer to the output of random numbers each time a subject arrives.<sup>i</sup>

<sup>i</sup>Be sure to make a notation on the paper questionnaire or use another method to keep track of which group, treatment, or control the subject was in if it is not obvious.

FIGURE 7.1        GRAPHPAD



Scientific Software    Data Analysis

# QuickCalcs

1. Select category

2. Choose calculator

3. Enter data


4. View results

Randomly assign subjects to treatment groups

Randomly choose a group for each subject

Assign  subjects to each of  groups. Repeat  times.

Do it!



# QuickCalcs

1. Select category

2. Choose calculator

Assign subjects to groups

Subject #	Group Assigned
1	A
2	A
3	B
4	B
5	B
6	A
7	B
8	B
9	A
10	A
11	B
12	A
13	A
14	B
15	A
16	A
17	B
18	A
19	B
20	A
21	B
22	B
23	A
24	B
25	A
26	A
27	B
28	A
29	A
30	A

Source: <http://www.graphpad.com/quickcalcs/randomize1.cfm>



FIGURE 7.2 ■ RANDOM ASSIGNMENT GENERATOR

Random Assignment Generator for a Factorial Experiment with Many Conditions

<p><b>Description</b></p> <p>This applet provides a list of random numbers that can be used to assign participants to conditions in an experiment with many conditions, such as a factorial experiment. (Read more about factorial experiments.)</p> <p>Enter the number of participants <math>N</math>, and the number of conditions <math>C</math>, for the experiment you are planning. This applet will then provide a random number for each participant. This will be a number from 1 to <math>C</math>. For example, if the 4th number in the list is 7, the 4th subject is randomly assigned to experiment condition 7. Random numbers will be generated so that the experiment is approximately balanced.</p>	<p><b>Generate list of random numbers for assignment</b></p> <p>Number of Participants, <math>N</math>: <input type="text" value="100"/></p> <p>Number of Conditions, <math>C</math>: <input type="text" value="2"/></p> <p><input type="button" value="Compute"/></p>
--	--

Random Selection

[Return to Sample Size Calculator](#)

[Download: CSV](#) | [Excel](#)

Number of Participants,  $N$  : 100

Number of Condition,  $C$ : 2

For Review-No Commercial Use(2023)

Assignments:

- 2
- 1
- 1
- 2
- 2
- 1
- 2
- 1
- 1
- 2
- 1
- 2
- 2
- 2
- 1
- 1
- 2
- 2
- 1
- 2
- 1
- 1
- 2
- 2
- 1
- 1

Online randomizers will automatically assign equal numbers of subjects to each group if the maximum number of subjects is divisible by the number of groups; for example, for three groups with forty subjects in each, set the total number of subjects at 120, not 125.

## Survey Experiments

One of the newest techniques revolutionizing the way experimental subjects are randomly assigned is with software designed to administer surveys online, such as Qualtrics and SurveyMonkey, among others. These **survey experiments**, also frequently described as an “experiment embedded in a survey,”<sup>23</sup> use a randomizer function to randomly assign subjects to treatment and control conditions without the researcher having to do anything other than set the randomizer before launching the survey. Computer-Assisted Telephone Interviewing software can also randomly assign subjects to conditions, similar to the survey software, with subjects taking the study over the phone instead of online.

Some researchers reserve the term *survey experiment* for studies that randomly sample from a population and also randomly assign subjects to conditions,<sup>24</sup> while others use the term when random sampling is not used<sup>25</sup> but random assignment is. Erisen<sup>26</sup> and colleagues make a distinction between survey experiments and lab experiments, explaining that lab experiments are conducted in a controlled environment where factors such as room temperature, time of day, or other things that could contaminate results can be controlled, whereas survey experiments can be taken by subjects in their own homes and with external factors out of the control of researchers, such as whether the subject takes a break to answer the door, go to the bathroom, or take a phone call. This is especially important when measuring things like the time a subject takes to complete the study, knowledge questions where subjects could look up the answers online,<sup>27</sup> or group decision studies where subjects may not believe they are interacting with real people over the Internet.<sup>28</sup> In cases where this kind of internal validity needs to be assured, survey software can be used with subjects in a lab, providing researchers with both control and the convenience of random assignment and data collection by computer. In a study of politicians’ decision making,<sup>29</sup> the researchers had subjects do the study online using survey software, but in order to maintain control, they performed the study in the presence of a researcher. Other researchers have explored ways to discourage cheating on knowledge questions by looking up answers so that online experiments are less prone to this kind of error.<sup>30</sup>

There has been much debate and research on the subject of whether survey experiments that use random samples are better because they are generalizable, or if convenience samples produce similar results. Many studies show little to no differences.<sup>31</sup> That issue was covered extensively in this book in the section on external validity in chapter 5. The point in this chapter is that whether it is called a survey experiment or lab experiment, the defining feature of all true experiments is that subjects must be randomly assigned to treatment and control groups.

## In-Person Randomizing

Having subjects participate in an experiment via computer makes random assignment easier on many levels. But sometimes an experiment is best conducted with paper and pencil. In-person studies require different random assignment strategies. For example, political scientists might conduct an experiment using voters as they exit polling stations to ensure that subjects actually voted. If a message in a pamphlet or newspaper is the message being tested, then using a printed version is more realistic than one shown online. Most studies on moral judgment are conducted in person because of the complexity of the topic.<sup>32</sup> Moral judgment experiments with journalists may be conducted in the newsrooms with the researcher providing lunch and having subjects take the study while they eat; this generates more participation from busy working professionals.<sup>33</sup> Other experiments may be conducted at professional association conferences, giving the study with paper and pencil at tables in the lobby.<sup>34</sup> For some studies, an old-fashioned paper-and-pencil test conducted in a classroom or library might provide faster data collection.<sup>35</sup> In these cases, random assignment needs to be conducted by online randomizers or old-fashioned methods described earlier.

## REPORTING RANDOM ASSIGNMENT

For Review-No Commercial Use(2023)

Regardless of how random assignment is achieved, it should always be reported in the research paper as having been done, and preferably the procedure used. No great amount of detail on the specifics is usually necessary, but the report should at least say that subjects were randomly assigned to conditions lest readers think otherwise. Here are two actual published examples:

“Participants were randomly assigned to a condition.”<sup>36</sup>

“Participants were then randomly given a booklet, which contained instructions for the research, a test pamphlet, and a response questionnaire. The test pamphlet was one of the four versions of a health pamphlet about HPV and genital warts.

Thus participants were randomly assigned to one of the experimental conditions.”<sup>37</sup>

The best practice is to provide a description of the randomization mechanism—for example, whether it was done with a deck of cards or die, or the randomizer function in a particular piece of software—and whether it was pregenerated or produced on site.<sup>38</sup>

## BALANCED AND UNBALANCED DESIGNS

Random assignment can raise concerns about unequal numbers of subjects in groups, as it did for me in my first experiment. This is called an **unbalanced design**. Very unequal sample sizes can affect group equivalence. It is not crucial to have exactly the same

number of subjects in each group; as long as the numbers in each group are close, it will be approximately balanced.<sup>39</sup> However, balanced designs, where the exact same numbers of subjects are in each group, give a study more statistical power (the subject of chapter 8).

There may be very practical and unavoidable reasons for an unbalanced design, such as some subjects dropping out or being purged. For example, in moral judgment studies, there is a built-in check for subjects who are trying to fake a better score.<sup>40</sup> When faking is detected, these subjects are eliminated from the data set, usually resulting in unequal numbers of subjects in groups. Sometimes, researchers lose subjects who did not complete enough questions for their data to be useful, or when subjects systematically drop out of a study, which is called **attrition**. Concerns about unequal numbers of subjects in groups should never be a reason to deviate from randomization, as failing to properly randomize is a far greater threat than unbalanced groups. As was pointed out in the discussion of control groups in chapter 6, having up to a third fewer subjects in the control group of an incomplete factorial is acceptable.<sup>41</sup> When treatments are expensive or difficult to run, having fewer subjects in the groups that are of less interest is also acceptable. In addition, when a treatment is desirable, people may be reluctant to participate if they think they may be denied it by being assigned to the control group. Having more subjects randomly assigned to the treatment group can help overcome these objections.<sup>42</sup>

It is also important to report attrition rates, as was done in a business experiment of a mentoring program. In that study, 52% of subjects dropped out of the study; the researcher compared the dropouts to those who remained in the study and found no statistical differences on the observable characteristics.<sup>43</sup> Rules of thumb for attrition rates say between 5% and 20% may be a source of bias.<sup>44</sup>

Many good statistics books explain ways of analyzing data from unbalanced designs. Statistical techniques such as Levene's test<sup>45</sup> can be used to determine if the unequal number of subjects results in unequal variance. When the Levene's test is significant, indicating the variance is not homogeneous, the researcher then uses more stringent tests of differences that do not assume equal variances. This is a subject for a statistics class or text. Suffice it to say that there are better ways to deal with unbalanced groups than by going off course with the randomization mechanism.

## CHECKING THAT RANDOM ASSIGNMENT WAS EFFECTIVE

---

Even though random assignment is the best method researchers have for getting equivalent groups, and randomization failure is rare,<sup>46</sup> some are skeptical. That is when a random assignment or balance check, which is a comparison of the groups' equivalence, is

in order. It is not necessary for *all* variables to be equally distributed among groups; those that are highly related to the outcome or dependent variables are of most concern. A rule of thumb for when something is “highly related” is that if a variable correlates with the dependent variable at .45 or greater, that variable must be equivalently distributed among groups.<sup>47</sup> It is a common misunderstanding that random assignment must result in equivalence on every variable known to mankind; that is not the case.<sup>ii</sup>

## Aggregate Level Random Assignment

Random assignment checks are more important when assignment is done at aggregate levels rather than with individuals; for example, businesses, families, polling precincts, or classrooms can be randomly assigned as units.<sup>48</sup> Aggregate level assignment<sup>iii</sup> is usually done when it is difficult or impossible to randomly assign individuals, but this is less preferable than individual random assignment, as people within a group or organization may differ systematically. For example, a start-up business may have younger workers than an established firm. An intact classroom may have group dynamics that create different motivation levels. Polling precincts may have voters who vary systematically because of the neighborhood they can afford to live in. It is more important to test equivalence when groups, such as classes or whole schools rather than individuals, are randomly assigned. This is because the sample size is usually lower when using aggregate groups than when using individuals.<sup>49</sup> For example, one study<sup>50</sup> assigned forty-one schools to three conditions, giving each group thirteen to fourteen schools. This is well below the suggested size of twenty.<sup>51</sup>

## Reporting Random Assignment Results

Not all journal articles will report the results of a random assignment check, but the practice is growing more common, although not without controversy.<sup>iv</sup> The field of political science has developed guidelines that require reporting whether random assignment was employed, as well as the unit that was randomized, and tables or text showing baseline means and standard

<sup>ii</sup>For a detailed discussion of this, see Mutz and Pemantle.

<sup>iii</sup>It is also important that the unit of *assignment* be the same as the unit of *analysis* in statistical tests. For example, if schools or classrooms are assigned as aggregate units to receive a treatment, then statistical analysis should be based on the aggregate level, not the individual level. In a hypothetical study of 2,500 students at thirty schools, fifteen of which are assigned to treatment and fifteen to control conditions, the total sample size is thirty, not 2,500. Otherwise, the precision of results will be overstated due to the overinflated N.

<sup>iv</sup>To get a sense of the controversy, read Gerber et al., (2014), the challenge to it by Mutz and Pemantle (2016), and Gerber et al.’s response (2016): Alan Gerber et al., “Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee,” *Journal of Experimental Political Science* 1, no. 1 (2014): 81–98; Alan S. Gerber et al., “Reporting Balance Tables, Response Rates and Manipulation Checks in Experimental Research: A Reply from the Committee That Prepared the Reporting Guidelines,” *ibid.* 2, no. 2 (2016): 216–29; Diana C. Mutz and Robin Pemantle, “Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods,” *ibid.*

deviations for certain variables.<sup>52</sup> To monitor the results of random assignment, researchers can test the equivalence of the groups on important variables with means and standard deviations or statistics designed to detect differences such as *t* tests, chi-square, and analysis of variance (ANOVA). These are then reported in a table. For example, if gender is important to the outcome variable, it will be checked to see that the groups have equivalent numbers of men and women; this is one time where finding no significant difference is a good thing. Only basic information is necessary for most randomization reports. For example, here is how one study reported it in an experiment testing the vividness effect on health messages:

Two-way ANOVA crossing the two manipulated variables (“message vividness” and “argument strength”) were performed on participants’ gender, age, sex behavior, and so forth. Results showed that the participants in the four experimental conditions were not significantly different from each other ( $p > .05$ ). Therefore, randomization appears to be effective.<sup>53</sup>

Some authors go further. For example, one study<sup>54</sup> included a table showing the descriptive statistics for various independent variables by group but no significance tests (see figure 7.3). Here is the narrative that was included in that study, and also the table:

### For Review-No Commercial Use(2023)

The means of the independent variables in each of the experimental conditions are reported in table 1. As shown in table 1, there were no substantial differences between the means and standard deviations of all the independent variables. The gender proportions in both conditions were also equivalent. Since the independent variables were asked before the manipulation, this shows that the random assignment indeed resulted in equal groups.<sup>55</sup>

Ho and McLeod included this table illustrating that random assignment resulted in equivalent groups on various variables.

Here is another example from a study on tutoring that did include the results of significance tests of random assignment and also included a table:

In order to test whether the students were distributed randomly in terms of these background measures, group means were calculated and *t*-tests were run to test for significant differences between the treatment and control groups. Tests of significance indicated that, for all background measures but one, there were no statistically significant differences between the treatment and control groups: the randomisation had worked to create pre-treatment equivalence.<sup>56</sup>

It is important to report randomization checks after a study has been completed if there has been considerable attrition, or subjects dropping out. This is to ensure that the dropouts did

**FIGURE 7.3** ■ TABLE OF RANDOM ASSIGNMENT

	FTF ( <i>n</i> = 192)		CMC ( <i>n</i> = 160)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Gender	Females (71.9%)	—	Females (69.1%)	—
Print news use	4.73	2.07	4.49	2.11
Television news use	5.02	1.96	4.98	2.04
Fear of isolation	2.98	1.10	3.26	1.10
Communication apprehension	3.00	1.23	3.03	1.26
Current opinion congruency	.09	.24	.09	.25
Future opinion congruency	.32	.50	.30	.49

From: Ho, Shirley S., and Douglas M. McLeod. 2008. "Social-Psychological Influences on Opinion Expression in Face-to-Face and Computer-Mediated Communication." *Communication Research* 35 (2) (April): 190–207.

## For Review-No Commercial Use(2023)

not differ significantly from the subjects who stayed in the study. Sometimes, the treatment itself is the reason. It may be too long, boring, or difficult, so subjects quit. Experiments that are conducted over a long period of time are especially prone to this problem.

### When Random Assignment Fails

Random assignment is the best available method for achieving equivalence among experimental groups, but it does not guarantee that groups will be perfectly matched on every individual difference variable.<sup>57</sup> It minimizes, rather than prevents, confounding.<sup>58</sup> Differences still may occur due to chance. It is not necessary to have equivalent groups on every possible variable; it is most important on variables that are correlated with the outcome variable.<sup>59</sup> For example, in moral judgment studies, age and education are highly correlated with moral judgment, but gender is not.<sup>60</sup> So it is more important to have groups be equivalent on age and education, and not worry so much about having equivalent numbers of men and women. When groups are not equivalent on important characteristics, internal validity decreases and researchers run the risk of making erroneous inferences.<sup>61</sup> However, failure to achieve equivalence with a proper randomization mechanism is rare.<sup>62</sup> Furthermore, the alpha level of significance testing already takes into account the fact that some variables will be spread unevenly across groups due to chance.<sup>63</sup> There is no way to “fix” a true failure of random assignment other than to start from scratch and redo the randomization properly.<sup>64</sup>

**FIGURE 7.4**  **RANDOM ASSIGNMENT REPORTING**

	FTF ( <i>n</i> = 192)		CMC ( <i>n</i> = 160)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Gender	Females (71.9%)	—	Females (69.1%)	—
Print news use	4.73	2.07	4.49	2.11
Television news use	5.02	1.96	4.98	2.04
Fear of isolation	2.98	1.10	3.26	1.10
Communication apprehension	3.00	1.23	3.03	1.26
Current opinion congruency	.09	.24	.09	.25
Future opinion congruency	.32	.50	.30	.49

From: Ritter, Gary W., and Rebecca A. Maynard. 2008. "Using the Right Design to Get the 'Wrong' Answer? Results of a Random Assignment Evaluation of a Volunteer Tutoring Programme." *Journal of Children's Services* 3 (2): 4–16.

## For Review-No Commercial Use(2023)

Here are some tips for achieving equivalence:

- If it is possible to include all subjects in all the groups without carry-over effects—that is, using a within-subjects design—then equivalence is a nonissue because the exact same people are in each group.
- Groups are considered nonequivalent if twice as many subjects in one group have some nuisance variable compared to the other group<sup>65</sup>—for example, if there are twice as many men as women in one group where gender is related to the outcome; ten men and five women in a group is considered nonequivalent.
- Start by ensuring there are enough subjects in each group. A power analysis, explained in chapter 8, will do this. The more subjects, the greater the chance that equivalence will be achieved.<sup>66</sup> In studies that employ multiple runs, more subjects can be recruited and randomly assigned, and the experiment conducted again to reach equivalence.<sup>67</sup>
- A good rule of thumb for achieving equivalence is to have at least twenty subjects in each group.<sup>68</sup> But even then, groups with fewer than twenty subjects actually are protected from erroneous conclusions of nonequivalence because statistics tests have a harder time detecting spurious differences with small numbers.<sup>69</sup>
- The random assignment process can be redone until equivalence is achieved.<sup>70</sup> Obviously, this only works if random assignment can be checked before the treatments are given.



- If nonequivalence can only be detected after the study is conducted, one widespread strategy is to use the nonequivalent variable as a covariate in the statistical analysis using analysis of covariance.<sup>71</sup> This will help reduce the influence of that variable before testing differences between the groups. For example, if there are twice as many men in the treatment group as the control group, and gender is expected to affect the outcome, using gender as a covariate will help level the playing field. That is, it helps remove unexplained variability due to the effects of gender before testing the effects of the treatment, which leads to more precise estimates.<sup>72</sup> This approach should be used conservatively, however, as covariates should only be used if planned in advance (a topic of the next chapter), as this will not “control” for the lack of true random assignment. And the growing tendency to include numerous covariates defeats the purpose of a well-designed, controlled experiment.<sup>73</sup>

One thing researchers should *not* do is purposively recruit more subjects with the needed characteristic and assign them just to the group low on that variable.<sup>74</sup> That is not random. Nor should a researcher try to rebalance the groups by moving subjects around to even out the groups before giving the treatment.<sup>75</sup> That also is not random.

Finally, it is somewhat reassuring to know that if groups are not equivalent on some variables, the differences will represent random error, not systematic error, and are unlikely to produce incorrect inferences.<sup>76</sup> Additionally, replication helps correct for any erroneous conclusions from a study threatened by nonequivalence; for more on that, refer to chapter 5. Over multiple studies, the truth tends to prevail.<sup>77</sup>

In fields where equivalence can be prone to problems, such as in education, social work, criminology, and program evaluations, a significant amount of time may need to be devoted to achieving equivalent groups. For example, in a study of a drug program in schools,<sup>78</sup> after random assignment was completed and checked for equivalence, two schools dropped out, making the groups nonequivalent. The researchers had to draw new schools, randomly assign them, and recheck equivalence.

## BLOCKING, MATCHING, AND OTHER STRATEGIES

One way to reduce the chances of unequal groups is with a **matched pairs** strategy and **blocking**. This involves matching subjects on important variables and then assigning them to treatment and control groups as a pair or block.<sup>79</sup> This is used to help ensure that an extraneous or nuisance<sup>80</sup> variable related to the outcome variable does not confound the results. Groups are created based on subjects with the same level of the blocking variable. For example, if gender is the blocking variable, subjects would be randomly paired by gender—a man with a woman—and then each pair randomly assigned to the treatment or control group.<sup>81</sup>

In a business study on the effectiveness of a mentoring program, the researcher was not allowed to randomly assign subjects to condition by the employers, who wanted to select employees with the highest potential for promotion to the program.<sup>82</sup> Instead, the study used a matched pairs design, selecting control group employees who were similar to treatment group employees based on five characteristics such as similarity of salary, performance rating, tenure in the organization, working in the same office, and not previously participating in a mentoring program. The study reports statistical tests showing no differences between treatment and control group subjects on these variables, while also noting “the treatment and matched control groups may have varied on unobserved characteristics.”<sup>83</sup>

Experimenters must anticipate and be able to measure the variable before conducting the study, so matching is no help for confounding variables that are not known; simple random assignment is still preferable for this reason.

Blocking also can be done with more than two levels of variables. For example, if age is the variable one wanted to ensure is equivalent across groups, then blocks of different age groups could be created; for example, four blocks for the age groups eighteen to thirty-four years old, thirty-five to fifty-four years old, fifty-five to sixty-four years old, and sixty-five years and older. After these blocks are created, the subjects in each are randomly assigned to treatment and control groups so that an equal number of subjects from each age block are in each group. Now, age cannot be the cause of any differences between the experimental groups.

Blocking is not preferable to straightforward random assignment but is useful when very few potential subjects are available or small sample sizes are likely—for example, when groups such as schools are the units. Matching strategies are frequently used in education studies, where schools are matched on important characteristics and then randomly assigned one to each condition.<sup>84</sup> Another example comes from a study on the effects of having grown up with friends of a different ethnicity on stereotyping.<sup>85</sup> The researcher anticipated it would be hard to find subjects who had grown up in integrated neighborhoods, so he pretested experimental volunteers, measuring their level of personal contact with minorities, and then matched high-contact subjects with low-contact subjects before randomly assigning them to the treatment or control group.

Blocking and matching strategies also can be useful when random assignment is not likely to be implemented correctly—for example, in program evaluations where the researcher is not in control of assignment.<sup>86</sup>

## Blocking vs. Simple Random Assignment

Matching strategies are *not* preferable to simple randomization,<sup>87</sup> although they can be misleading in their intuitive appeal.<sup>88</sup> For one thing, statistical tests lose power when blocking factors do not have much influence on the outcome variable.<sup>89</sup> Blocking is

more common in the life sciences when studying plants and animals.<sup>90</sup> Another drawback is that blocking requires a two-step process, first measuring subjects on the blocking factor, then randomly assigning them to groups, administering the treatment, and measuring the outcome. Social science experts agree that simple random assignment is preferable to other methods for achieving comparable groups.<sup>91</sup> Even the harshest critics of random assignment do not argue for an alternative.<sup>92</sup> Campbell and Stanley are particularly critical of the “widespread and mistaken preference . . . for equation through matching,”<sup>93</sup> saying, “matching is no real help when used to overcome initial group differences.” When confounding variables are unknown, and so uncontrollable (sometimes called *lurking variables*), random assignment is the best strategy, as it automatically balances these.<sup>94</sup>

## Stratified Random Assignment

One technique that makes it easier to assess equivalence on many variables is **stratified random assignment**, where numerous variables are combined into a single variable, similar to a factor created by factor analysis.<sup>95</sup> This does not block or match up subjects or units using any particular variables, but rather combines numerous related variables into one overarching factor that can be measured for equivalence after subjects are randomly assigned. In other words, it allows researchers to assess equivalence on multiple variables instead of worrying about numerous discrete variables. For example, one study<sup>96</sup> used seven key variables such as the type of community a school was in, number of grades in the school, percentage of White students, enrollment per grade, and rural or urban setting. It then used a statistical technique to arrive at one composite stratifying variable. This was done to find a combination of variables that were closely related within these schools, which the authors called “rurality,” explaining that rural schools often were similar on these characteristics. After random assignment, equivalence was checked by the single composite factor rather than on seven individual variables. The actual process was more complicated than described here.<sup>97</sup> The authors found equivalence using ANOVA tests of differences and reported it in a table (see figure 7.6). This stratifying procedure has the advantage of having unknown or unmeasured variables be randomly distributed across groups, whereas blocking and matching strategies do not.<sup>98</sup> As with achieving equivalence, this strategy is reserved for variables that are likely to be highly correlated with the outcome variable, not for every variable.<sup>99</sup> This technique, like blocking, is more common in some disciplines than others, so it is important to know the standard in your field.

## RANDOM ASSIGNMENT OF OTHER THINGS

---

So far, this chapter has focused on randomly assigning individual subjects to conditions. But random assignment applies to more than simply how subjects are assigned to groups. In fact, experts advise randomly assigning as many of a study’s procedures as possible.<sup>101</sup>

**FIGURE 7.5**          **RANDOM ASSIGNMENT CHECK**

The text explaining the random assignment equivalence check said, “First, we verified that the assignment procedure worked to provide group equivalence on the school-level variables relating to CIS score. We conducted a simple ANOVA (SAS Proc GLM), with Program (Rural, Classic, Control) listed as a class variable. Table 3 presents the results for the overall F test. As expected, program group membership was unrelated to the CIS score, the two factors making up the CIS score, and the seven items making up the factors.”<sup>100</sup>

	Tutored students	Control students	Total sample
Average reading grade, 1997-98 year-end (A = 4.0)	1.65	1.61	1.63
Average math grade, 1997-98 year-end (A = 4.0)	1.79	1.85	1.82
Average SAT-9 reading open-ended national percentile score, 1998	23.1	23.2	23.2
Average SAT-9 math open-ended national percentile score, 1998	16.7	18.6	17.6
Full year attendance rate, 1997-98	91.8	90.6	91.2
% of students not promoted, 1997-98	9.2	9.0	9.1
% of students African American	95.9	96.8	96.4
% in home receiving welfare assistance (AFDC)	63.3	62.9	62.1
% with guardian with a high school degree or more	68.6	74.5	71.5
% with guardian currently working for pay	57.2	51.9	54.6
% with guardian reporting health problem that limits activity	19.5	23.2	21.3
% in home with both mother and father	37.4	31.1	34.4
% reporting that someone helps with homework	73.2	78.4	75.7
% reporting that someone at home reads with them	67.0	69.5	68.2
Average number of children in household	3.39	3.02	3.20
<b>Total study sample</b>	<b>196</b>	<b>189</b>	<b>385</b>

1. The welfare assistance data were derived from the School District of Philadelphia student information system.
2. The background data related to student guardians and the numbers of children per household were derived from the baseline survey completed by the guardians who gave parental consent for the student to participate in the tutoring programme (September 1998).
3. The figures on household composition and help with reading and homework were derived from the tutee follow-up survey administered to tutees at programme completion in May 1999.

Shadish, Cook, and Campbell deliberately refer to random assignment of “units” so as not to imply that only people should be randomized.<sup>102</sup> Anything that could introduce systematic bias should be randomly assigned. For example, if more than one experimenter

**FIGURE 7.6** ■ TESTING FOR RANDOM ASSIGNMENT TABLE

Variable	<i>F</i> (2,38)	<i>p</i>
CIS	0.25	.78
Factor 1	1.05	.36
Factor 2	0.08	.93
Rurality	0.45	.64
Npergrade	1.62	.21
Numgrades	0.41	.67
Pctwhite	1.60	.21
Pctlunch	0.10	.90
Scores	0.22	.81
Rdrugs	0.80	.46

Note: Post hoc tests with the Duncan test showed no significant differences ( $p < .10$ ) for any individual comparisons.

will administer the study, the experimenters should be randomly assigned to the sessions and conditions he or she will supervise.<sup>103</sup> An experimenter who observes and measures subjects might become better after practice or, conversely, grow tired and get worse at measuring. It is important that the experimenter not be assigned to observe all subjects in the control group first and the treatment group last (or vice versa) in order to avoid introducing systematic differences between the groups. Instead, experimenters should be randomly assigned to each group as well as to session.

Typically, when there is more than one stimulus—advertisements, news stories, and health messages, for example—these should be presented to subjects in a random order. In a study of politicians' decision making, the researchers randomized many things.<sup>104</sup> In addition to randomly assigning subjects to condition, they randomized the three types of tasks and also the thirteen decisions within each type of task that subjects had to make. The study also gave subjects an incentive in the form of a donation to a charity, and the way that was offered was also randomized; here is the explanation:

To make each decision that includes a monetary payoff relevant, but simultaneously ensure that tasks do not influence each other, we randomly selected one task that determined how much money was donated to the charity on behalf of the participant. Specifically, we randomly selected either the lottery-choice or the lottery-valuation part of the experiment, and then we randomly selected one task from this part. This avoided that participants' choices were influenced by so-called portfolio effects (e.g., some safe and some risky choices for a balanced portfolio) or by previous earnings.<sup>105</sup>

This study used a within-subjects design, but because the scenarios in the two conditions (gain frames, loss frames) were similar, the researchers did not want subjects to read each scenario in both conditions, so they randomly assigned each participant to read half the scenarios in each condition. They explain it like this:

For each scenario, it was randomly determined whether a participant was presented with the loss or the gain frame, so it was possible that a participant would get the gain frame for one scenario and a loss frame for another scenario. We also randomly determined the order of the scenarios.<sup>106</sup>

Clearly, these researchers followed Bausell's advice: "When in doubt—randomize."<sup>107</sup>

## Counterbalancing

The reasons for randomly assigning or rotating the order of something, or **counterbalancing**, is to avoid carry-over effects. These were described in chapter 6 as effects due to learning, practice, fatigue, or the subjects changing. Carry-over effects happen when receiving one treatment affects a subject's response to the next treatment. One specific kind of carry-over effect arises from the order in which things are presented, known as **order effects**. Order effects have been well documented under the study of **primacy and recency**, the ideas that we remember best what we are exposed to first and last. This is a special concern for within-subjects designs where every subject gets all the different treatments, as they are especially prone to fatigue, practice effects, carry-over, and order effects.<sup>108</sup> Counterbalancing is helpful because often the carry-over effect in one direction will cancel out the effect in the other direction. For example, some subjects may do better on the last treatment (recency effect) and others on the first (primacy effect). When the data are aggregated, these two effects cancel each other out. The same applies to everything that is randomly assigned; for example, if more than one experimenter will supervise multiple runs of a study, the different experimenters should not only be randomly assigned to the treatment or control conditions, they should also be rotated, or counterbalanced, among sessions.<sup>109</sup> The goal of counterbalancing is to make order effects equivalent across the conditions. And, as with all things equivalent, failing to counterbalance orders decreases internal validity.

## Latin Square

Counterbalancing can be accomplished by simply randomly assigning, but there is one specific type of counterbalancing strategy used in experimental research called the **Latin square** that equalizes the number of positions under which each stimulus occurs. It ensures that each experimental message or stimulus occurs in the first position one time, in the last position one time, and in each in-between position one time. Also, each condition or stimulus follows the other exactly once. This is more efficient than simple random assignment.<sup>110</sup> Here is an illustration of how it works: A hypothetical experiment uses

three different advertisements as stimuli. If the researcher randomly assigns the ads to order, there are six possible order combinations:

A	B	C
A	C	B
B	C	A
C	B	A
C	A	B
B	A	C

But using Latin squares produces three combinations:

A	B	C
B	C	A
C	A	B

Each advertisement is shown first one time, last one time, and in the middle one time. The two key features of a Latin square are that a row or column never contains the same letter twice and that every row and column contains the same letters. This is much more efficient than simple random assignment to order,<sup>111</sup> as it reduces the number of subjects needed to receive each order.<sup>112</sup> Latin squares are especially useful in large factorial designs where it would be quite costly to administer all possible order combinations.<sup>113</sup> This type of counterbalancing is achieved by randomly selecting stimuli to represent A, B, and C in the first row.<sup>114</sup> Next, the stimuli are simply rotated by moving the first-place stimulus to last place in each row and sliding the others over one place. This works for three or more stimuli; obviously with two stimuli, there are only two possible order combinations.

In reporting Latin squares, as in reporting random assignment, it is common to just see it mentioned in passing. For example, in this study of public service announcements (PSAs) in a health study, the authors said, “The order of the presentation of PSAs was counterbalanced according to a diagram-balanced Latin square.”<sup>115</sup>

And in another study of decision making in TV newsrooms, the author described the rotation of the three story scenarios this way: “The order of which story participants received was counterbalanced using a Latin squares design.”<sup>116</sup>

In studies that use multiple factors, Latin squares help avoid confounding of the factors. For example, in a study of the use of humor in advertising,<sup>117</sup> the authors had subjects listen to radio broadcasts that featured different advertisements. The factors used to make up each ad consisted of the type of product (e.g., cereal, cheese, batteries), brand name, and jokes, which they called one-liners. Here is how they describe their counterbalancing strategy:

Rotations for the three versions of radio broadcasts were designed such that a particular one-liner was not presented with a particular brand name or product type more than once in the study. . . . Within each combination, product types, brand names, and one-liners were rotated in order or presentation for each of the three audiotapes. To arrange three commentaries in three rotations, a Latin square counterbalancing technique (Keppel, 1991) was used.<sup>118</sup>

The Latin squares design got its name from an ancient puzzle on the different ways Latin letters could be arranged in a square.<sup>119</sup> It was introduced as a rotation experiment,<sup>120</sup> popularized by R. A. Fisher,<sup>121</sup> and has been the preferred method in psychology ever since.<sup>122</sup>

## RANDOM ASSIGNMENT RESISTANCE

---

Most researchers are quickly persuaded of the abilities of randomization to solve a multitude of problems, but that is frequently not the case for nonresearchers whose participation may be needed in a study. Disciplines that perform program evaluations or conduct studies in real-world or field environments may encounter resistance to random assignment. For example, school administrators allowed some students to bypass the random assignment process in one study.<sup>123</sup> Education researchers, for example, find that school personnel may object to offering some students a treatment opportunity and denying it to others.<sup>124</sup> Business researchers have found that executives may refuse to allow their employees to be randomly assigned, instead insisting on hand-picking those assigned to each group themselves.<sup>125</sup> In a health study, “a substantial number” of subjects refused to participate in the randomization because they did not want to be involved if they were not assured of receiving the potentially beneficial treatment.<sup>126</sup> Ethical concerns also raise opposition to randomization; for example, in an experiment on treatment for crime victims, some subjects who exhibited self-harming tendencies were moved from the control group to the treatment group, disrupting the randomization.<sup>127</sup> This can lead to researchers abandoning a true experiment in favor of a quasi experiment, discussed in chapter 4. This topic is discussed in more detail in More About box 7.3.

As elegant a procedure as random assignment is in all its forms, it is not perfect. But as Campbell and Stanley say, “It is nonetheless the only way of doing so, and the essential way.”<sup>180</sup>

With a firm understanding of how subjects (and other things) should be assigned, the next chapter will deal with ways to sample subjects and how to determine the optimum number of them.



## MORE ABOUT . . . BOX 7.3

### Resistance to Random Assignment

When it comes to random assignment, compliance is critical.<sup>128</sup> As with pregnancy, there is no such thing as being “a little bit random”<sup>129</sup> — it is an all-or-nothing proposition. That distinction is often lost on nonresearchers; for example, in one study, when told that random assignment had been compromised, staff implementing it acted surprised and responded that they believed the assignments were “in fact, more-or-less random.”<sup>130</sup> One employee directly involved with the process believed the instructions were merely “recommendations.”<sup>131</sup>

Another example of random assignment gone wrong includes a murder by a subject who dropped out of a program, which led the prosecutor to refuse to deny treatment to anyone in need.<sup>132</sup>

Not all threats to random assignment are this dramatic. Any actions that compromise randomization can undermine the internal validity of an experiment by leading to nonrandom differences between subjects in treatment and control conditions. In the following table, the first column lists some common objections to random assignment from practitioners in the field tasked with assigning subjects to conditions for researchers. The column on the right contains advice for overcoming these objections. Researchers who have reported their results with these techniques have seen compliance with randomization go from as low as 19% to 94%.<sup>133</sup>

The advice here is drawn from education, counseling, criminal justice, and business fields but applies to attempts to assess the effectiveness of an intervention in a field setting in any discipline.

The best approach is for researchers to insist on conducting random assignment themselves and to carry out the procedure at the researcher’s site, not the study site.<sup>134</sup> This should always be done in conjunction with communication with the site staff, allowing them to voice concerns, have questions answered, and participate in the design process.<sup>135</sup> Monitoring the assignment process is also essential. Subjects have been known to try to sneak into groups to which they are not assigned.<sup>136</sup> It is also important to observe the extent to which the intervention is actually being implemented; for example, if teachers are supposed to use technology, check to see how much they are doing so.<sup>137</sup>

OBJECTION	RESPONSE
<b>FAIRNESS</b>	
Giving a perceived benefit, even if unproven, to some and not all is intrinsically unfair. <sup>138</sup>	Explain that when resources are limited and there are more people who need services than slots available, random assignment is one of the fairest ways to distribute services. <sup>140</sup>
School personnel, in particular, are not inclined to offer interventions to some and deny others the same opportunity. <sup>139</sup>	Point out that random assignment protects the organization from accusations of favoritism. <sup>141</sup>
	Explain the random assignment process as a lottery, which gives everyone an equal chance, which is fair. <sup>142</sup>
	Alter the control group so they receive a lower dose of the treatment instead of no treatment. <sup>143</sup>
	Agree to give those in the control group priority to participate in the treatment group during another run of the experiment. <sup>144</sup>

(Continued)

(Continued)

OBJECTION	RESPONSE
<p>NEED</p> <p>Fears that the most in need of the treatment would not be chosen by random assignment.<sup>145</sup></p>	<p>Offer to categorize some of the most in need as “wildcards” who can bypass the random assignment process and are put in the treatment group; then exclude these subjects from the analysis.<sup>146</sup></p> <p>Divert the most needy into an alternative intervention that is not part of the study.<sup>147</sup></p> <p>Explain that the treatment has not yet been shown to work—that is what the study is for. If it does not provide benefit, then the needy in the control group will not have lost anything.<sup>148</sup></p> <p>Explain that if the treatment is shown to work in this randomized experiment, everyone can be offered it later.<sup>149</sup></p>
<p>EVALUATION FEARS</p> <p>If the experiment shows the program has no impact, it could be eliminated.<sup>150</sup></p>	<p>Be sensitive to this issue and collaborate on the outcome variables. Add in qualitative data that can provide insights beyond the quantitative. Also include measures that may show more sensitivity to the program.<sup>151</sup></p>
<p>CONFLICTS WITH PRACTICE</p> <p>Randomly assigning people to receive no training or be enrolled in certain classes conflicts with normal procedures.<sup>152</sup></p> <p>It may be difficult to assign certain interventions to only parts of existing units. For example, in schools where classes are taught by teams of teachers, assigning an intervention to only one set of teachers can lead to group planning issues.<sup>153</sup></p> <p>In one study, special education students needed to be in the same class, resulting in special education students being overenrolled in one condition.<sup>154</sup></p> <p>Staff is uncomfortable with an outsider telling them what to do.<sup>155</sup></p> <p>It might send unintended signals to high-performing employees if they were not chosen for the treatment.<sup>156</sup></p> <p>Staff may object to the extra work of implementing random assignment, experience scheduling conflicts, or key people may leave due to illness or turnover.<sup>157</sup></p>	<p>Get to know staff, the environment, and their needs.<sup>158</sup></p> <p>Suggest procedures that allow for a minimum of disruption.<sup>159</sup></p> <p>Allow practitioners to exempt up to 10% of subjects from random assignment for practical reasons. Track them and exclude from the analysis.<sup>160</sup></p>

OBJECTION	RESPONSE
<p><b>ACCIDENTAL</b></p> <p>Subjects may be accidentally assigned to the wrong condition due to misunderstanding, lack of time to make assignments properly, or staff turnover.<sup>161</sup></p> <p>In one study of schools, staff did not understand the lists of students already included random assignment to group, and they devised their own method to randomly assign from the lists.<sup>162</sup></p>	<p>Improve communication.<sup>163</sup></p> <p>Assign a randomization liaison from the research team to the organization.<sup>164</sup></p> <p>Schedule multiple meetings and information sessions to explain the research design.<sup>165</sup></p> <p>Include everyone. For example, in one study, security guards were in charge of checking subjects into the program, so they needed to know why it was important that subjects go to the group they were assigned.<sup>166</sup></p> <p>Include staff in initial planning sessions for random assignment.<sup>167</sup></p> <p>Look for staff and organizations that have prior experience with random assignment and research studies. In one study, a staff member involved with random assignment said it took a year and a half before she fully understood the assignment process.<sup>168</sup></p> <p>Use color-coded forms to make it easier to quickly see to which condition a subject should be assigned. A study of domestic violence treatment trained police officers in random assignment using color-coded report forms.<sup>169</sup></p>
<p><b>SUBJECTS OBJECT</b></p> <p>Subjects request being put in different groups, refuse to participate in the randomization, withdraw in the middle of treatment, or complain to administrators, who reassign them.<sup>170</sup></p> <p>Subjects drop out for a variety of reasons including health, lack of childcare, and transportation issues.<sup>171</sup></p>	<p>Make the control group more attractive by giving subjects some sort of “treatment as usual” such as normal classroom instruction or other interesting activities that will not confound the outcome.<sup>172</sup> One set of researchers considered an open-ended discussion group as an enhancement to the basic treatment.<sup>173</sup></p> <p>Establish a procedure so that those requesting to be moved must meet with the researchers, who explain the reasons behind the assignment.<sup>174</sup></p> <p>Offer the intervention to the subject at a later time.<sup>175</sup></p> <p>Put control group subjects who wish to receive the intervention on a waiting list for future programs.<sup>176</sup></p> <p>In one study, a principal insisted that siblings of those chosen for the treatment group also had to be included. Exclude these subjects from statistical analysis.<sup>177</sup></p> <p>Have staff refer subjects’ requests to researchers, who are in a better position to turn them down and explain why.<sup>178</sup></p> <p>Track subjects who leave and learn why.<sup>179</sup></p>

For Review-No Commercial Use(2023)

## STUDY SPOTLIGHT 7.4

### Taking advantage of random assignment in a natural setting

**From: Abrams, David S., and Albert H. Yoon. 2007. "The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability." *University of Chicago Law Review* 74 (4) (Fall): 1145–1177.**

This study took advantage of naturally occurring random assignment when the researchers discovered that a county in Nevada was assigning incoming felony cases randomly to attorneys in the pool, which allowed for a natural experiment free from selection bias.

Clark County, which includes Las Vegas, began assigning attorneys to cases after a defendant's death sentence was overturned because he was assigned to an inexperienced public defender. Under the previous nonrandom assignment method, the better attorneys might be assigned the more difficult cases, thus confounding attorney ability with case difficulty. This opportunity allowed the researchers to examine the performance of attorneys that cannot be explained by case characteristics. Conventional wisdom says that lawyers who attend better law schools may get clients lower sentences, for example. The study discovered that Hispanic attorneys and those with more experience achieve better outcomes for clients than others, but gender and law school attended made no difference.

For the purposes of this book, the study is important because it illustrates in depth the value of random assignment.

The researchers began by checking to see that random assignment was indeed being implemented and not thwarted. This helped rule out alternative explanations for trial outcomes, such as case difficulty. Cases were assigned to attorneys without the judges, prosecutor, or team chief knowing any characteristics of the defendant or even the alleged offense, helping to ensure against any subconscious efforts to assign cases purposively. The researchers used nonparametric tests (chi-square) to see if cases were indeed being assigned randomly using the defendants' age, gender, and race. They explain that these three variables are highly correlated with other defendant characteristics on unobserved variables. They say, "Crucially, we assume that this provides evidence that unobservables are also randomly assigned [due to correlation with observables]."<sup>181</sup>

#### TESTING FOR RANDOM ASSIGNMENT

Case characteristic	<i>p</i> -value	Observations
Defendant sex	0.851	10,129
Defendant age	0.253	9,803
Defendant race	0.098	7,145

Note: Each row reports results from a separate simulation to test for the equality of public defender fixed effects. Defendant sex is a dummy variable for whether the defendant is male. Defendant race is 0 for black defendants and 1 for white defendants.

Nonsignificant results showed the cases were indeed randomly assigned (see figure 7.6). This then allowed the researchers to test their hypotheses concerning the differences among attorneys' abilities and other variables that predict better trial outcomes for defendants.

These researchers were creative in spotting an opportunity for a natural experiment. As with most journal articles, this one does not mention the researchers gaining approval from an Institutional Review Board (IRB) before collecting data. The data used may have been exempt from informed consent because it was a matter of public record, or the researchers may have given informed consent after the fact.

New researchers sometimes assume that because data are already collected, this constitutes "secondary data" and one does not need approval from an IRB. True secondary data occurs when subjects have received informed consent—for example, in existing data sets such as the American National Election Studies or Pew polls. The organizations collecting these data sets and making them available to researchers have obtained permission from a human subjects committee of an IRB and have provided subjects with informed consent. This is not the case with all existing data. For example, students who fill out evaluations about satisfaction with their classes have not been given the information contained in informed consent documents. In any research that involves information collected from human subjects, researchers should contact their IRB to determine if they need to obtain IRB approval and subjects' informed consent, even if data have already been collected. Chapter 11 will discuss informed consent and IRB approval in more detail. Researchers should be attuned to opportunities for naturally occurring randomization but should also be careful to secure permission from an IRB and follow protocols for obtaining informed consent from the people whose data will be used.

For Review-No Commercial Use(2023)

## Common Mistakes

- Not randomly assigning subjects to groups, or not doing random assignment appropriately
- Failing to randomly assign other elements of a study, such as the experimenters, to sessions
- Not counterbalancing stimuli

## Test Your Knowledge

1. When a participant has an equal chance of being in the treatment or control group in an experiment, it is called \_\_\_\_\_.
  - a. Random sampling
  - b. Random assignment

- c. Random error
  - d. Selection bias
2. The main reason for using random assignment in an experiment is to ensure which of the following?
    - a. A sample representative of the population
    - b. That neither subject nor experimenter knows which group someone is in
    - c. That groups are as equivalent as possible on known and unknown variables
    - d. That the dependent variable does not differ across conditions
  3. Which of the following is NOT a way to randomly assign subjects to groups?
    - a. Drawing names out of a hat
    - b. Flipping a coin
    - c. Using a random number generator
    - d. Rotating subjects so that groups come out even (e.g., 1, 2, 1, 2 . . .)
  4. Groups need to be equivalent on all variables that can be measured.
    - a. True
    - b. False
  5. Which is the preferred way to ensure that systematic variation does not confound a study?
    - a. Pretesting
    - b. Blocking or matching
    - c. Simple random assignment
    - d. Balancing groups by moving subjects around after random assignment
  6. What besides subjects should be randomly assigned?
    - a. Nothing, only the subjects
    - b. Experimenters
    - c. All of the study's procedures
    - d. All but A
  7. Latin square is a technique for \_\_\_\_\_.
    - a. Creating equivalent groups
    - b. Controlling for extraneous individual variables
    - c. Minimizing order effects
    - d. Randomly assigning subjects to groups

For Review-No Commercial Use(2023)

8. Complete the following to make a Latin square:

A	B	C	D

9. If men and women are paired and then assigned to the treatment or control group as a pair, this is called \_\_\_\_.
- a. Blocking or matching
  - b. Random assignment
  - c. Stratified random assignment
  - d. Latin square
10. One drawback to random assignment is that experimenters must anticipate and be able to measure confounding variables before conducting the study, so it is no help for confounding variables that are not known.
- a. True
  - b. False

For Review-No Commercial Use(2023)

Answers

- 1. b
- 2. c
- 3. d
- 4. b

- 5. c
- 6. d
- 7. c

8.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

- 9. a
- 10. b

Application Exercises

- 1. Using the experimental study you began developing in chapter 1 and continued by creating a design table and control group for in chapter 5, decide how you will randomly assign subjects to groups. Write one page on which strategy you will use and why—random number generator, drawing out of a hat, etc. Do a “test run” of this. Assume forty subjects in each of the groups in your study. Use your choice of randomizer to assign subjects. Analyze the results; were they balanced or unbalanced? Repeat the

process with a different type of randomizer to see how it turns out (i.e., if you used an online randomizer first, repeat by drawing out of a hat).

2. Write one page about all the elements of your study that could be randomly assigned. Besides the subjects, what other procedures might be randomly assigned? Why? Assume you will use at least three stimuli in your study (e.g., three different treatments, interventions, teaching techniques, ads, PSAs, stories, messages, etc.). Design a plan for counterbalancing them.
3. Write one page about the possible confounding variables that random assignment will need to help equalize across groups. Read literature about your outcome variable to see what others have found to be highly correlated with your dependent variable. Use your imagination and common sense to identify as many as you can.

## Suggested Readings

Chapter 5 in: R. Barker Bausell. 1994. *Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist*. Thousand Oaks, CA: Sage.

Chapter 8, “Randomized Experiments: Rationale, Designs, and Conditions Conducive to Doing Them,” pp. 246–278, in: Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth Cengage Learning.

Read these two articles, one critical of random assignment and one more supportive, for a comparative perspective:

- Krause, Merton S., and Kenneth I. Howard. 2003. “What Random Assignment Does and Does Not Do.” *Journal of Clinical Psychology* 59 (7): 751–766.
- Strube, M. J. 1991. “Small Sample Failure of Random Assignment: A Further Examination.” *Journal of Consulting & Clinical Psychology* 59 (2): 346–350.

## Notes

1. W. A. McCall, *How to Experiment in Education* (New York: MacMillan, 1923), 41.
2. R. Barker Bausell, *Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist* (Thousand Oaks, CA: Sage, 1994).
3. Richard A. Berk, “Randomized Experiments as the Bronze Standard,” *Journal of Experimental Criminology* 1 (2005): 416–433.
4. Graeme D. Ruxton and Nick Colegrave, *Experimental Design for the Life Sciences*, 3rd ed. (Oxford, UK: Oxford University Press, 2011).
5. Bausell, *Conducting Meaningful Experiments*.
6. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963).
7. Charles Sanders Peirce and Joseph Jastrow, “On Small Differences of Sensation,” *Memoirs of the National Academy of Sciences* 3, no. 75–83 (1885): 8.
8. William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Belmont, CA: Wadsworth Cengage Learning, 2002).



9. J. M. Gueron, "The Politics of Random Assignment: Implementing Studies and Impacting Policy," in *Evidence Matters: Randomized Trials in Education Research*, ed. F. Mosteller and R. Boruch (Washington, DC: Brookings Institution Press, 2002), 26.
10. Diana C. Mutz and Robin Pemantle, "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods," *Journal of Experimental Political Science* 2, no. 2 (2016): 192–215.
11. Martin T. Orne, "Demand Characteristics and the Concept of Quasi Controls," in *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*, ed. Robert Rosenthal and Ralph L. Rosnow (Oxford: Oxford University Press, 2009), 110–137.
12. Cathaleene Macias et al., "Preference in Random Assignment: Implications for the Interpretation of Randomized Trials," *Administration & Policy in Mental Health & Mental Health Services Research* 36, no. 5 (2009): 331–342.
13. T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Chicago: Rand-McNally, 1979).
14. Colin Ong-Dean, Carolyn Huie Hofstetter, and Betsy R. Strick, "Challenges and Dilemmas in Implementing Random Assignment in Educational Research," *American Journal of Evaluation* 32, no. 1 (2011): 40.
15. Macias et al., "Preference in Random Assignment," 331–342.
16. Ibid.
17. D. F. Halpern, *Thought and Knowledge: An Introduction to Critical Thinking*, 4th ed. (Mahwah, NJ: Erlbaum, 2003); M. Patricia King et al., "Impact of Participant and Physician Intervention Preferences on Randomized Trials: A Systematic Review," *Journal of the American Medical Association* 293, no. 9 (2005): 1089–1099.
18. Macias et al., "Preference in Random Assignment," 331–342.
19. Ibid.
20. Ibid.
21. Sameer B. Srivastava, "Network Intervention: Assessing the Effects of Formal Mentoring on Workplace Networks," *Social Forces* 94, no. 1 (2015): 427–452.
22. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
23. Cengiz Erisen, Elif Erisen, and Binnur Ozkececi-Taner, "Research Methods in Political Psychology," *Turkish Studies* 13, no. 1 (2013): 17.
24. Diana Mutz, *Population-Based Survey Experiments* (Princeton, NJ: Princeton University Press, 2011); Steven L. Nock and Thomas M. Guterbock, "Survey Experiments," in *Handbook of Survey Research*, ed. P. V. Marsden and J. D. Wright (Bingley, UK: Emerald Group Publishing Limited, 2010), 837–864.
25. Erin C. Cassese et al., "Socially Mediated Internet Surveys: Recruiting Participants for Online Experiments," *PS: Political Science & Politics* 46, no. 4 (2013): 1–10; James N. Druckman, "Priming the Vote: Campaign Effects in a US Senate Election," *Political Psychology* 25, no. 4 (2004): 577–594; Julio J. Elias, Nicola Lacetera, and Mario Macis, "Markets and Morals: An Experimental Survey Study," *PLoS ONE* 10, no. 6 (2015): 1–13.
26. Erisen, Erisen, and Ozkececi-Taner, "Research Methods in Political Psychology," 17.
27. Scott Clifford and Jennifer Jerit, "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Computer and Online Studies," *Journal of Experimental Political Science* 1, no. 2 (2014): 120–131.
28. Rebecca B. Morton and Kenneth C. Williams, *Experimental Political Science and the Study of Causality: From Nature to the Lab* (New York: Cambridge University Press, 2010).
29. Jona Linde and Barbara Vis, "Do Politicians Take Risks Like the Rest of Us? An Experimental Test of Prospect Theory under MPs," *Political Psychology* 38, no. 1 (2017): 101–117.
30. Clifford and Jerit, "Is There a Cost to Convenience?" 120–131.
31. Cassese et al., "Socially Mediated Internet Surveys"; Kevin J. Mullinix et al., "The Generalizability of Survey Experiments," *Journal of Experimental Political Science* 2 (2015): 109–138.
32. Renita Coleman and Lee Wilkins, "The Moral Development of Journalists: A Comparison With Other Professions and a Model for Predicting High Quality Ethical Reasoning," *Journalism and Mass Communication Quarterly* 81, no. 3 (2004): 511–527.
33. Renita Coleman and Lee Wilkins, "Searching for the Ethical Journalist: An Exploratory Study of the Ethical Development of News Workers," *Journal of Mass Media*

- Ethics* 17, no. 3 (2002): 209–255; Renita Coleman and Lee Wilkins, “The Moral Development of Journalists”; R. Coleman and L. Wilkins, “The Moral Development of Public Relations Practitioners: A Comparison with Other Professions,” *Journal of Public Relations Research* 21, no. 3 (2009): 318–340.
34. Renita Coleman, “The Moral Judgment of Minority Journalists: Evidence from Asian American, Black, and Hispanic Professional Journalists,” *Mass Communication and Society* 14, no. 5 (2011): 578–599.
  35. Erisen, Erisen, and Ozkececi-Taner, “Research Methods in Political Psychology,” 17.
  36. Kate E. West, “Who Is Making the Decisions? A Study of Television Journalists, Their Bosses, and Consultant-Based Market Research,” *Journal of Broadcasting & Electronic Media* 55, no. 1 (2011): 27.
  37. Jun Myers, “Stalking the ‘Vividness Effect’ in the Preventive Health Message: The Moderating Role of Argument Quality on the Effectiveness of Message Vividness,” *Journal of Promotion Management* 20, no. 5 (2014): 634.
  38. Mutz and Pemantle, “Standards for Experimental Research,” 192–215.
  39. Bausell, *Conducting Meaningful Experiments*.
  40. Mark L. Davison and Stephen Robbins, “The Reliability and Validity of Objective Indices of Moral Development,” *Applied Psychological Measurement* 2, no. 3 (1978): 391–403; James R. Rest, *Development in Judging Moral Issues* (Minneapolis, MN: University of Minnesota Press, 1979).
  41. Bausell, *Conducting Meaningful Experiments*.
  42. Christopher H. Rhoads and Charles Dye, “Optimal Design for Two-Level Random Assignment and Regression Discontinuity Studies,” *Journal of Experimental Education* 84, no. 3 (2016): 421–448.
  43. Srivastava, “Network Intervention.”
  44. D. Fergusson et al., “Post-Randomisation Exclusions: The Intention to Treat Principle and Excluding Patients from Analysis,” *BMJ* 325 (2002): 652–654.
  45. Howard Levene, “Robust Tests for Equality of Variances,” in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. Ingram Olkin et al. (Stanford, CA: Stanford University Press, 1960), 278–292.
  46. Mutz and Pemantle, “Standards for Experimental Research,” 192–215.
  47. M. J. Strube, “Small Sample Failure of Random Assignment: A Further Examination,” *Journal of Consulting & Clinical Psychology* 59, no. 2 (1991): 346–350.
  48. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
  49. John Graham et al., “Random Assignment of Schools to Groups in the Drug Resistance Strategies Rural Project: Some New Methodological Twists,” *Prevention Science* 15, no. 4 (2014): 516–525.
  50. Ibid.
  51. L. M. Hsu, “Random Sampling, Randomization, and Equivalence of Contrasted Groups in Psychotherapy Outcome Research,” *Journal of Consulting and Clinical Psychology* 57 (1989): 131–137.
  52. Alan Gerber et al., “Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee,” *ibid.* 1, no. 1 (2014): 81–98.
  53. Myers, “Stalking the ‘Vividness Effect,’” 636.
  54. Shirley S. Ho and Douglas M. McLeod, “Social-Psychological Influences on Opinion Expression in Face-to-Face and Computer-Mediated Communication,” *Communication Research* 35, no. 2 (April 2008): 190–207.
  55. Ibid., 197.
  56. Gary W. Ritter and Rebecca A. Maynard, “Using the Right Design to Get the ‘Wrong’ Answer? Results of a Random Assignment Evaluation of a Volunteer Tutoring Programme,” *Journal of Children’s Services* 3, no. 2 (2008): 4–16.
  57. Merton S. Krause and Kenneth I. Howard, “What Random Assignment Does and Does Not Do,” *Journal of Clinical Psychology* 59, no. 7 (2003): 751–766.
  58. Ibid.
  59. Bausell, *Conducting Meaningful Experiments*; Hsu, “Random Sampling”; Strube, “Small Sample Failure of Random Assignment,” 346–350.
  60. James R. Rest et al., *Postconventional Moral Thinking: A Neo-Kohlbergian Approach* (Mahwah, NJ: Erlbaum, 1999).
  61. Krause and Howard, “What Random Assignment Does and Does Not Do.”
  62. Mutz and Pemantle, “Standards for Experimental Research.”
  63. Ibid.
  64. Ibid.

65. Hsu, "Random Sampling."
66. Strube, "Small Sample Failure of Random Assignment."
67. Graham et al., "Random Assignment of Schools."
68. Hsu, "Random Sampling."
69. Strube, "Small Sample Failure of Random Assignment."
70. Bausell, *Conducting Meaningful Experiments*.
71. Ibid.
72. Donald P. Green and Daniel Winik, "Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism among Drug Offenders," *Criminology* 48, no. 2 (2010): 357–387.
73. Mutz and Pemantle, "Standards for Experimental Research."
74. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
75. Ibid.
76. Strube, "Small Sample Failure of Random Assignment," 346–350.
77. Krause and Howard, "What Random Assignment Does and Does Not Do"; Strube, "Small Sample Failure of Random Assignment."
78. Graham et al., "Random Assignment of Schools."
79. Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*.
80. Murray Webster and Jane Sell, *Laboratory Experiments in the Social Sciences* (Amsterdam: Elsevier, 2007).
81. Bausell, *Conducting Meaningful Experiments*; Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*.
82. Srivastava, "Network Intervention."
83. Ibid., 438.
84. Graham et al., "Random Assignment of Schools."
85. David Alan Free, "Perpetuating Stereotypes in Television News: The Influence of Interracial Contact on Content" (unpublished dissertation, University of Texas, 2012).
86. E. W. Gondolf, "Lessons from a Successful and Failed Random Assignment Testing Batterer Program Innovation," *Journal of Experimental Criminology* 6 (2010): 355–376.
87. Mutz and Pemantle, "Standards for Experimental Research."
88. Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*.
89. Ruxton and Colegrave, *Experimental Design for the Life Sciences*.
90. For example, see *ibid.*; Murray R. Selwyn, *Principles of Experimental Design for the Life Sciences* (Boca Raton, FL: CRC Press, 1996).
91. Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*; Strube, "Small Sample Failure of Random Assignment."
92. Krause and Howard, "What Random Assignment Does and Does Not Do."
93. Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*, 15.
94. Selwyn, *Principles of Experimental Design for the Life Sciences*.
95. Graham et al., "Random Assignment of Schools."
96. Ibid.
97. Ibid.
98. Ibid.
99. Bausell, *Conducting Meaningful Experiments*.
100. Graham et al., "Random Assignment of Schools," 521.
101. Ibid.
102. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 253.
103. Bausell, *Conducting Meaningful Experiments*.
104. Linde and Vis, "Do Politicians Take Risks Like the Rest of Us?"
105. Ibid., 107–108.
106. Ibid., 111.
107. Bausell, *Conducting Meaningful Experiments*, 78.
108. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
109. Bausell, *Conducting Meaningful Experiments*.
110. Roger Kirk, *Experimental Design: Procedures for the Behavioral Sciences*, 4th ed. (Los Angeles, CA: Sage, 2013).
111. Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*.
112. Ruxton and Colegrave, *Experimental Design for the Life Sciences*.
113. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*.
114. Kirk, *Experimental Design*.
115. Zhang Jueman, Zhang Di, and T. Makana Chock, "Effects of HIV/AIDS Public Service Announcements on Attitude and Behavior: Interplay of Perceived Threat

- and Self-Efficacy," *Social Behavior and Personality: An International Journal* 42, no. 5 (2014): 799–809.
116. West, "Who Is Making the Decisions?" 27.
  117. Eron M. Berg and Louis G. Lippman, "Does Humor in Radio Advertising Affect Recognition of Novel Product Brand Names?" *Journal of General Psychology* 128, no. 2 (2001): 194.
  118. Ibid., 199.
  119. Kirk, *Experimental Design*, 671.
  120. E. L. Thorndike, W. A. McCall, and J. C. Chapman, *Ventilation in Relation to Mental Work*, vol. 78, Teachers College Contribution to Education (New York: Teachers College, Columbia University, 1916).
  121. Ronald A. Fisher, *Statistical Methods for Research Workers* (Edinburgh: Oliver and Boyd, 1925).
  122. Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*.
  123. Ritter and Maynard, "Using the Right Design to Get the 'Wrong' Answer?"
  124. Ibid.
  125. Srivastava, "Network Intervention."
  126. Yun Hyung Koog and Byung-Il Min, "Does Random Participant Assignment Cause Fewer Effects in Research Participants? Systematic Review of Partially Randomized Acupuncture Trials," *Journal of Alternative and Complementary Medicine* 15, no. 10 (2009): 1107–1113.
  127. Richard A. Berk, Gordon K. Smyth, and Lawrence W. Sherman, "When Random Assignment Fails: Some Lessons from the Minneapolis Spouse Abuse Experiment," *Journal of Quantitative Criminology* 4, no. 3 (1988): 209–223.
  128. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas."
  129. Gueron, "The Politics of Random Assignment," 26.
  130. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas," 39.
  131. Ibid., 45.
  132. Gondolf, "Lessons from a Successful and Failed Random Assignment."
  133. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas."
  134. Gary W. Ritter and Marc J. Holley, "Lessons for Conducting Random Assignment in Schools," *Journal of Children's Services* 3, no. 2 (2008): 28–39.
  135. Ibid.
  136. Ibid.
  137. Ibid.
  138. Ibid.
  139. Ritter and Maynard, "Using the Right Design."
  140. Ritter and Holley, "Lessons for Conducting Random Assignment in Schools."
  141. Ibid.
  142. Bausell, *Conducting Meaningful Experiments*; Ritter and Holley, "Lessons for Conducting Random Assignment in Schools."
  143. Ibid.
  144. Ibid.
  145. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas"; Ritter and Holley, "Lessons for Conducting Random Assignment in Schools."
  146. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas"; Ritter and Holley, "Lessons for Conducting Random Assignment in Schools."
  147. Berk, "Randomized Experiments as the Bronze Standard"; Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas."
  148. Bausell, *Conducting Meaningful Experiments*.
  149. Ibid.
  150. Ritter and Holley, "Lessons for Conducting Random Assignment in Schools."
  151. Ibid.
  152. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas."
  153. Ritter and Holley, "Lessons for Conducting Random Assignment in Schools."
  154. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas."
  155. Ibid.
  156. Srivastava, "Network Intervention."
  157. Gondolf, "Lessons from a Successful and Failed Random Assignment."
  158. Ritter and Holley, "Lessons for Conducting Random Assignment in Schools."
  159. Bausell, *Conducting Meaningful Experiments*.
  160. Ong-Dean, Huie Hofstetter, and Strick, "Challenges and Dilemmas."
  161. Ibid.
  162. Ibid.
  163. Ibid.
  164. Ibid.

165. Ritter and Holley, “Lessons for Conducting Random Assignment in Schools.”
166. Ibid.
167. Ong-Dean, Huie Hofstetter, and Strick, “Challenges and Dilemmas.”
168. Ibid.; Ritter and Holley, “Lessons for Conducting Random Assignment in Schools.”
169. Berk, Smyth, and Sherman, “When Random Assignment Fails.”
170. Ong-Dean, Huie Hofstetter, and Strick, “Challenges and Dilemmas”; Koog and Min.
171. Berk, Smyth, and Sherman, “When Random Assignment Fails.”
172. Bausell, *Conducting Meaningful Experiments*, 74.
173. Gondolf, “Lessons from a Successful and Failed Random Assignment.”
174. Ong-Dean, Huie Hofstetter, and Strick, “Challenges and Dilemmas.”
175. Ibid.
176. Ritter and Holley, “Lessons for Conducting Random Assignment in Schools.”
177. Ibid.
178. Ibid.
179. Ibid.
180. Campbell and Stanley, *Experimental and Quasi-Experimental Designs for Research*, 15.
181. David S. Abrams and Albert H. Yoon, “The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability,” *University of Chicago Law Review* 74, no. 4 (2007): 1145–1177.

For Review-No Commercial Use(2023)



For Review-No Commercial Use(2023)



## SAMPLING AND EFFECT SIZES

*When some would have us eliminate college students as a source of information we could ask, tongue-in-cheek, “What is it that occurs when a person’s hand touches their degree that suddenly makes him or her a valid subject in a . . . study where they would not*

*have been a minute earlier?”<sup>1</sup>*  
**For Review-No Commercial Use(2023)**

—Michael Basil

*Our findings would be substantially more credible if students were not so often the first and only choice.<sup>2</sup>*

—William Wells

### LEARNING OBJECTIVES

- Think critically about the use of students and subjects from various sources.
- Create a plan for recruiting and incentivizing subjects for an experiment.
- Describe the relationships between power and sample size.
- Explain effect sizes.
- Prepare and report the results of a power analysis.

There is nothing more important to experiments than the topic of the last chapter—random assignment. It may seem as if this book has put the cart before the horse in that it discusses how subjects are assigned before it addresses subjects themselves. The reason for this is that understanding random assignment is key to recognizing why certain subjects are appropriate for experiments when they may not be for other methods. This chapter is concerned with the primary focus of what is being randomly assigned—that is, the subjects in the sample. Exploring the related issues of power, which helps determine how many subjects are needed and effect sizes, or how well the treatment worked, is another aim of this chapter.

This chapter builds upon the discussion of external validity from chapter 5 while drawing from the understanding of random assignment in chapter 7. To briefly recap, the type of subjects that are used in an experiment can affect how well the results generalize to others. One of the most serious criticisms of experiments is that they usually cannot be generalized beyond the specific people in the study.<sup>3</sup> As was pointed out in chapter 5, most experiments use **convenience samples**—that is, people who are easily available—rather than those who are randomly sampled from a population.<sup>4</sup> Rather than random *sampling*, the counterpart in experiments is random *assignment*, which helps assure that subjects in different groups are equivalent on key characteristics. Equivalence is far more important in experiments than is the ability to generalize. Experiments then rely on logical inference to extend causal findings beyond the subjects of the study.

Experiments' superiority on internal validity and the benefits of random assignment do not give researchers a "pass" to use any-old subjects that are easily available. While we do refer to them as "convenience samples," the researchers' convenience is not the only thing that matters when selecting subjects. Researchers should carefully think through their choice of subjects for every study and then explain the reasons in the methods section of resulting papers. This chapter will supply fodder for those thoughts and explanations, and give an overview of some of the common sources of samples in experimental research. Assuming readers have now become reasonably assured that random sampling is, in most cases, an "inadequate" way to collect subjects for an experiment,<sup>5</sup> and that the use of nonrepresentative convenience samples in experiments is justified,<sup>6</sup> this chapter turns to one of the most often used and most frequently criticized type of subjects—students.

## STUDENT SAMPLES

---

There is no more convenient subject for a researcher in academia than a college student. Studies have shown that between 75% and 90% of experimental subjects in marketing and communication journals are students.<sup>7</sup> The tax discipline runs around 52%.<sup>8</sup>



Some departments make it easy to use student subjects by providing **subject pools** or **participant pools** where students are required to sign up to take part in a certain number of research studies per semester for course credit. This is common in courses with a research focus, such as methods classes, and in fields like psychology and advertising where students learn about the methods and procedures in practice firsthand. For example, in advertising, many of the decisions that go into determining what type of ad is produced result from advertising or market research, so it is logical that students should learn about the process that influences these choices. Students in subject pools are available to faculty researchers and sometimes also to graduate students. Even in departments where there is no pool of easily available student subjects, researchers find ways to entice students to participate in research with extra credit, gift cards, pizza, and the like. Yet one of the most common criticisms from reviewers is that a study is limited by the use of students as subjects. Some even feel that using students is “prescientific” or only valid in exploratory research and pilot studies.<sup>9</sup> Reviewers have been known to reject studies primarily on the basis of the use of student subjects. Before reaching that point, it behooves experimentalists to think carefully about the use of students—or anyone, for that matter—as subjects, choose the best subjects for the study, and then provide a strong theoretical rationale for the subjects used in the study. A sample of how researchers can do this is shown in the **How To Do It** box 8.1. What follows lays the groundwork for that thinking, decision making, and subsequent rationale.

Using students as samples can sometimes be problematic, not only for external validity in that they may not be generalizable but also for internal validity when their responses may be biased because they are students. Obviously, generalizability is not an issue where students are the population to be generalized to, such as education studies, but for others, problems with the use of students as subjects seems to fall into the following categories outlined by Sears.<sup>17</sup>

### Higher Than Average Cognitive Skills

Because college is a place designed to foster critical thinking, and also because taking tests and filling out questionnaires is a common activity, students may perform better on studies than people who have never been to college or who have been out for some time.<sup>18</sup> This may “change the nature of the mental processes that are performed,” says Basil,<sup>19</sup> who gives an example of research that investigates decision making as one area that may be different for students. The flip side is that students’ higher cognitive skills can actually lower random error because they can answer questions on studies and report their thoughts more accurately.<sup>20</sup> Researchers should be careful to assess whether cognitive functioning will bias a study; if there is no theory, evidence, or logical reason why it

## HOW TO DO IT 8.1

### Rationales for Subjects

Here are some examples of how researchers have described and justified their choice of samples.

#### College Students Examples

Here is the justification for use of college students from a political study that goes beyond the usual statements of how student samples are adequate and there is no reason to envision differences between students and nonstudents, and instead refers specifically to evidence on the topic of the study to support the claim:

"One criticism of this methodology (e.g., Graber, 2004) is that because, like many other media experiments (e.g., Druckman, 2001a; Miller & Krosnick, 2000; Nelson & Oxley, 1999), the present studies use college students as participants, the findings cannot be generalized to other populations. However, media effects such as agenda setting, priming, and framing have been replicated in experiments that use a variety of samples, from college students (e.g., Druckman, 2001a; Nelson & Oxley, 1999; Miller & Krosnick, 2000) to adult volunteers from cities in which the investigators live (e.g., Iyengar & Kinder, 1987) to the general population (e.g., Nelson & Kinder, 1996), and in studies that match content analyses of the media to surveys of the general population (e.g., Iyengar & Simon, 1993). Therefore, the criticism that using a college student sample will lead to qualitatively different conclusions than a general population sample is not supported by the evidence."<sup>10</sup>

For Review-No Commercial Use(2023)

Miller, J. M. 2007. "Examining the Mediators of Agenda Setting: A New Experimental Paradigm Reveals the Role of Emotions." *Political Psychology* 28 (6): 689–717.

Here is another one that goes well beyond the usual to give information specific to the study:

"Experiments generally require a well-defined participant population that can be tracked, a treatment that can be randomized and administered to the correct person, and the ability to measure the outcome of interest for both the treatment and control groups. Voter mobilization experiments fit these requirements by focusing on registered voters, of whom there is an official list that can be randomized and later updated with turnout. Unfortunately, an official list of unregistered persons does not exist. Even if such a list did exist, residential mobility is much higher among unregistered persons and the reliability of such a list would be suspect. Conducting the experiments on college campuses solves the problem with defining a participant population . . . .

College students are an interesting population to study with regard to voter registration for four reasons. First, college students are geographically mobile and are extremely likely to have moved in the recent past, necessitating re-registration (Squire, Wolfinger, and Glass 1987). This mobility can also remove students from social support networks (like parents) who can help students engage civically. Second, college students are young and less likely to have developed a habit of voting (Plutzer 2002; Bendor, Diermeier, and Ting 2003; Green, Green, and Shachar 2003; Fowler 2006). The flipside of habit formation is that gains in participation at a young age will translate into greater participation in the future. Third, college students fall into many of the demographic categories associated with low levels of electoral participation: young (Wolfinger and Rosenstone 1980), disinterested in politics (Verba, Schlozman, and Brady 1995), and news nonconsumers (Wattenberg 2007).

Finally, the federal government has mandated that colleges and universities make an effort to register students. A 1998 amendment to the Higher Education Act requires colleges and universities to distribute voter registration forms to students enrolled in all degree or certificate programs. Institutions

that fail to comply with the provision could jeopardize their federal student aid funds. Thus, college students present an empirically, pragmatically, and normatively interesting population to study with regard to registration habits.

College students are an especially interesting population to study with regard to e-mail and online registration tactics because they are more reliant on and frequent users of e-mail and the Internet relative to other age cohorts (Tedesco 2006). E-mail has not been found to be effective at boosting voter turnout (Nickerson 2007a), but college students should be the population most responsive to e-mail and Internet appeals. If e-mail messages encouraging registration and driving traffic to Web-based registration tools will work for any population, it would be college students."<sup>11</sup>

Bennion, E. A., and D. W. Nickerson. 2011. "The Cost of Convenience: An Experiment Showing E-Mail Outreach Decreases Voter Registration." *Political Research Quarterly* 64 (4): 858–869.

Here is another one that uses college students, justifying them because of their experience in the profession: "One-hundred-and-ninety-four journalism students at a large Midwestern university participated. Use of students is appropriate because this is designed to see if effects are present at all; it is not concerned with generalizability. The sample was limited to journalism students to approximate the thought processes of journalists.

Many students had significant professional experience before entering school. Once admitted, these students are required as part of their courses to work in the newsrooms of the newspaper, magazine, or network-affiliate television station which serve the entire community but are run by the school."<sup>12</sup>

Coleman, R. 2006. "The Effect of Visuals on Ethical Reasoning: What a Photograph Worth to Journalists Making Moral Decisions?" *Journalism and Mass Communication Quarterly* 83 (4): 835–850.

### MTurk Example

In this political communication study on belief echoes, or how false information can continue to influence a person's inferences even when corrected, the author used MTurk for subjects. Here is the reporting and justification:

"The experiments were conducted between August 2011 and November 2012. Participants accessed the experiment through Amazon's Mechanical Turk platform (MTurk).<sup>2</sup> Mechanical Turk is an online platform where subjects are paid to perform tasks ranging from captioning photos to taking surveys. The study was restricted to only United States participants over the age of 18.<sup>3</sup> Participants were paid between \$0.61 and \$0.75 for their participation (payment varied based on the length of the survey) and the surveys took most people between eight and 12 minutes to complete. All subjects were screened to ensure that none participated in more than one study. The three experiments described in this article each contain a unique set of participants.<sup>4</sup> A table describing the demographic characteristic of each sample is in the supplemental Appendix.

Several studies suggest that MTurk is a reasonable alternative to other traditionally used convenience samples.<sup>5</sup> While similar to the general population in many respects, the MTurk sample skews younger and more liberal. There is no theoretical reason to expect the formation of belief echoes to vary by age, as the cognitive processes that generate them should be similar among all ages. While partisanship might plausibly shape belief echoes, the experiments take this into consideration by randomly presenting subjects with misinformation that either reinforces or contradicts their preexisting partisan preferences. Each experiment followed a similar format. I explain this format in detail in the description of Experiment 1, then describe any variations in the descriptions of Experiments 2 and 3."<sup>13</sup>

(Continued)



SAGE Journal Article:  
study.sagepub.com/  
coleman

For Review No Commercial Use(2023)

(Continued)

Thorson, Emily. (2016). "Belief Echoes: The Persistent Effects of Corrected Misinformation." *Political Communication* 33 (3): 460–480.

### Panel Subjects Examples

In this study on the impact of government funding on donations to arts organizations, the authors use subjects from a firm that provides participants for research and marketing studies. Here is how they described it:

"Participants in the experiment were members of the CivicPanel project, an opt-in email panel of approximately 12,000 active panelists at the time of this study. CivicPanel is a university-affiliated online research project created to provide a general population of volunteers to participate in surveys and online studies about public and civic affairs. Voluntary panelists have been continuously recruited from various online postings including Google advertising, Craigslist.org, and the open directory Dmoz.org. . . . Given the substantive focus on the U.S. nonprofit sector, 133 responses from non-U.S. residents were dropped from the analysis. An additional 24 of those with partially completed responses were also removed, resulting in a final analytical sample  $n = 562$ .

Table 2 displays the number of participants in each of the four arms or treatment groups as well as their demographic characteristics and political attitudes. . . . In a typical full randomized setup, treatment, and control groups should have the same characteristics except for the treatment they are given (Remler & Van Ryzin, 2011). The lack of statistically significant differences across groups randomly assigned to each condition in Table 2 confirms the statistical equivalence of the experimental groups."<sup>14</sup>

Kim, M., and G. G. Van Ryzin. 2014. "Impact of Government Funding on Donations to Arts Organizations: A Survey Experiment." *Nonprofit and Voluntary Sector Quarterly* 43 (5): 910–925.

Here is another example of a study that uses a panel of participants:

"Participants were recruited from a census representative panel maintained by the Qualtrics research firm, which uses the stratified quota sampling method. In exchange for their participation, cash value rewards were credited to panelists' online accounts. A total of 2,301 adult citizens aged 18 and older were randomly selected from the panel and invited via e-mail to participate in an online study; 861 members agreed to participate, and 768 individuals successfully completed the study. This represents a cooperation rate of 33.4% [COOP1, AAPOR]."<sup>15</sup>

Lee, H., and N. Kwak. 2014. "The Affect Effect of Political Satire: Sarcastic Humor, Negative Emotions, and Political Participation." *Mass Communication and Society* 17 (3): 307–328.

### Professional Samples

Here is a study that used professionals as subjects:

"Participants were professional journalists at newspapers and television stations across the South and Southwest region of the United States. They were recruited by the researcher writing and telephoning managers for permission to come to their newsrooms to conduct a study. No managers refused to allow the researcher to come. Newsroom managers announced the study, its date, time, and location to employees in advance via email. The researcher then traveled to the newsrooms where participants completed the questionnaire, usually in a conference room, and the researcher provided either lunch on the premises or gift cards to local restaurants in exchange for participation. The entire study took approximately 30 minutes."<sup>16</sup>

Coleman, R. 2011. "Journalists' Moral Judgment About Children: Do As I Say, Not As I Do." *Journalism Practice* 5 (3): 257–271.

For Review No Commercial Use(2023)