the past—for example, how **randomization**, or the process of assigning subjects to groups, came about, and why experimentalists use terms like *split plots*. Some stories highlight ingenious ways to answer questions, encouraging researchers to think creatively about their own studies. Creativity is important in research not just for novelty's sake but also to help answer questions in the best way possible. Other examples show the importance of protecting participants from harm and how they helped lead to the development of Institutional Review Boards (IRBs), committees charged with ensuring researchers follow ethical guidelines. Many of these stories are quite entertaining, and some have even been made into movies for mass audiences. More important, understanding the historical roots can lead to a better grasp of contemporary practices. Other examples show how findings from experiments can affect real-world problems—for example, the famous studies that examined why good people do bad things. In this chapter are some of the stories of the "giants" from the quote by Isaac Newton upon whose shoulders experimental methodology is built.

The history of experimental design is filled with fascinating studies, often hyped as the most evil, creepy, bizarre, or ones that went horribly wrong. Studies that could never be conducted again are also a popular theme in the "mad science" category. This chapter will take a different approach; instead of dwelling on the findings or theories created by these experiments, it will highlight the development and early use of important discoveries such as random assignment, controls, and the use of confederates. Importantly, how ethical guidelines developed and IRBs came about will be included. (Chapter 11 will discuss IRBs and other ethical issues not fully addressed here in more detail.) Many of these studies are already known through popular folklore; in those cases, I try not to repeat the obvious but instead highlight aspects that are not as familiar. This chapter does not purport to be exhaustive or cover every important experiment ever conducted; of necessity, many are left out. In particular, the received history tends to leave out women and experimentalists of color. This chapter deals with some of these forgotten figures in More About . . . box 2.1, Contributions of Women. Many historians trace the development of experiments back to the ancient Greeks or others. After the introduction of one very early and important medical experiment, I devote the rest of this chapter to more modern social science experiments.

THE SCURVY STUDIES

Some of the earliest experiments were conducted on a disease called scurvy in 1747. Rare today, it was particularly problematic on ships as it caused sailors to become weak and anemic, and also caused their skin to bleed and gums to rot.[23] A ship's surgeon, James Lind, carried out one of the first controlled experiments to find a cure.[24] He chose twelve men who all had scurvy, using only men who "were as similar as I could have them."[25] He then divided them into six groups, putting two men in each. These six

MORE ABOUT . . . BOX 2.1
Contributions of Women

While men are most frequently cited as early pioneers of experimental design in the social sciences, there were also many important women, including women of color. These scientists faced barriers and discrimination in their careers, including a lack of fellowships and being barred from admission to graduate programs and employment in academic positions that allowed research and publishing.[1] Some women completed doctoral work, including theses and dissertations, but were denied the degrees. Men received credit for some of the contributions of women. Some women collaborated with their husbands, and antinepotism policies prevented them from being hired. Racial discrimination placed even more burdens on minority women.[2]

What few histories include women primarily note their theoretical contributions rather than their advancements in the methodology of experimental design, which is the subject of this book.[3] Next are short profiles of three women, including one African American, who made contributions to experimental methodology.[4] More certainly deserve recognition.

Mary Whiton Calkins is one of the pioneers of experimental psychology, remembered for, among other things, inventing the paired-associates task, a test of memory using paired numbers and colors.[5] Despite having completed all the requirements for a doctorate, Harvard refused to grant her a degree in 1890 because she was a woman.[6] Yale and the University of Michigan offered her admission, but she turned them down because they lacked a laboratory for experiments, which Harvard had. It was easier for women to find academic positions at women's colleges, so Calkins went to work at Wellesley and established an experimental laboratory there. Despite not having a doctorate, Calkins published four books and more than 100 papers in scientific journals.[7] Her accomplishments led to Columbia University and Smith College awarding her honorary doctorates. Despite this recognition and a petition to Harvard signed by thirteen graduates who were prestigious alumni, she was again denied a degree in 1927.[8]

William Notman

Calkins crossed paths with another experimentalist in this chapter, Joseph Jastrow, who along with Charles S. Peirce helped identify the benefits of random assignment. Jastrow published a study on the kinds of words produced by men and women when asked to write them out quickly, concluding that women's words were repetitive, individual, and concrete, whereas men's words were constructive, useful, and abstract.[9] She criticized his conclusion, pointing out the confounding effects of the environment, training, and socialization of women.[10] She was an advocate for women's rights all her life, repeatedly refusing to accept a doctoral degree from Radcliffe for work she did at Harvard.

New York Post Archives/Contributor

Mamie Phipps Clark was an African American psychologist noted for developing the Clark Doll Test for her research on race, which was used in the 1954 *Brown v. Board of Education* case that allowed African American students to attend White schools.[11] Unlike Calkins, Clark earned a doctoral degree in 1943, becoming the first African American woman to do so from

(Continued)

(Continued)

Columbia University. The first African American male to earn a doctorate from Columbia was her husband and research partner, Kenneth B. Clark.[12] He always credited her with the idea for the doll test.[13]

Mamie Clark's contribution to experimental methods involved showing children two identical dolls, one Black and one White. The children were asked which doll was bad and which was good, which one the child liked to play with, and which one most looked like them. Many Black children identified the Black doll as bad, and almost half said the White doll looked most like them. This was more pronounced in children from segregated schools than integrated ones.[14] The Clarks' experiments were important evidence that segregation harmed children, and also were influential in the first mass-produced doll of a Black infant.[15] Clark never found a position in academia, instead working as a researcher and then clinical psychologist at a children's home until she and her husband opened a testing and consultation center for minority children in Harlem.[16]

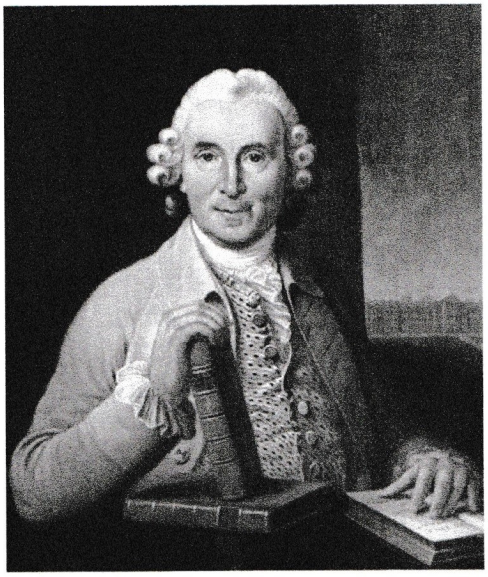
JHU Sheridan Libraries/Gado/Contributor

Mary Ainsworth is remembered for an assessment technique known as the "Strange Situation."[17] Work using this method helped advance psychologists' understanding of children's attachment to their caregivers. Ainsworth's method involved researchers observing through a one-way mirror a child's behavior during eight different episodes of about three minutes each where a mother and child come to the researcher's laboratory, which is filled with toys. The different episodes involve a stranger entering the room and trying to befriend the child, the mother leaving the child alone with the stranger, returning, both mother and stranger leaving the child alone, the stranger re-entering to comfort the child, and finally, the stranger picking the child up.[18] Observations were recorded every 15 seconds on a 1 to 7 scale. The assessment had good reliability, meaning other researchers could reproduce the findings.[19] Typical for a laboratory experiment, the method was criticized as being artificial and lacking in ecological validity,[20] and also on ethical grounds for causing stress in young children.[21]

Many other women also contributed to experimental methodology by authoring or coauthoring other tests and measures, including Grace Kent-Rosanoff's Word Association Test, Florence Goodenough's Draw-a-Man Test, and Grace Fernald's character tests that preceded other tests of moral development, among many others.[22]

groups then got different treatments ranging from a quart of cider; sulphuric acid; a half pint of seawater; a mixture of garlic, mustard, and horseradish in a dose "the bigness of a nutmeg"[26]; vinegar; or two oranges and a lemon.[27] Lind had six treatment conditions for this medical study, which is considered a lot in social science today, with two to four more common.[28] He also had a very small number of subjects—only twelve, with just two in each treatment group. Small sample sizes are more typical in medical studies, where large effects are more common than in social science. Another important issue in this study was the problem of how to assign people to conditions. Lind recognized the dilemma with groups where people were dissimilar on important factors, understanding that people's inherent individual differences, such as age, weight,

and general health, could affect the results. He limited his sample to men who were as similar as possible on certain requirements in order to get results of the treatment that were not confounded by these extraneous factors. Here, he used what would later be known as a **matching** strategy, pairing up men with similar characteristics. Importantly, he manipulated only these six conditions while holding everything else (that he could think of) constant. His independent variables—the six different dietary supplements—were not drawn from theory but from cures that had been proposed earlier. However, Lind did write about how the theory on scurvy was mostly conjecture by researchers who had never seen it, and advocated a mix of theory and hands-on experience. His measurement of one of the treatments' doses—the size of a nutmeg—is not very precise; today, that would be measured in grams or something similar.



Wikimedia Commons

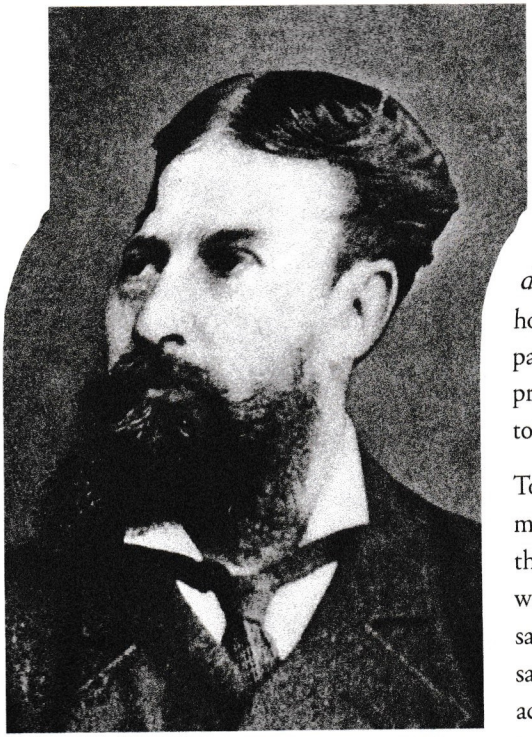
James Lind

Lind's study was a huge success by any standard. After only six days, the men who were given the oranges and lemon recovered.[29] The other men also improved, but the ones who ate citrus fruit had a dramatic recovery compared to the others, which eventually led to the recognition that vitamin C was the agent at work. No statistical tests of significance were performed, but the effects at six days were obvious.

This study was important for several reasons; it recognized that people needed to be assigned to treatment groups in such a way that individual differences would not matter, which Lind accomplished by carefully selecting men who had similar characteristics. This was in 1747. By the late 1800s, assigning subjects to conditions had become a burning topic and a different solution emerged, although it was applied to the treatments rather than the subjects, as illustrated in the next story.

THE CONTRIBUTIONS OF CHARLES PEIRCE

Charles Sanders Peirce and his student Joseph Jastrow were by some accounts the first to use random assignment.[30] The first study to use it was conducted in 1885 to see if people could judge how much something weighed just by feeling and looking at it.[31] The theoretical construct Peirce (pronounced "purse") was interested in was the source of judgment errors, resulting in concepts such as the just-noticeable difference, which business, marketing, and advertising scholars will recognize. In the first study, Peirce began the first experiment by always starting and ending with the heaviest weights. In the second study, he tried alternating the heaviest with the lightest weight. Last, he describes how he used a pack of cards



www.britannica.com

Charles Sanders Peirce

to randomly assign the order of the weights. His results were vastly different from the first two trials when the order of the weights had not been randomized, showing the importance of this technique.[32] Peirce used a deck of cards and simply shuffled them before drawing out a card that would determine which weight would come next. *Today, we tend to use random number generators, but a deck of cards will still work just as well. He talked about* how this method occasionally produced "long runs of one particular kind of change, which would occasionally be produced by chance," but notes that this was preferable to the subject knowing there would be no such patterns.[33]

To most people, the idea of something being "random" means it is unpredictable; but to Peirce, random meant that "in the long run any one individual of the whole lot would get taken as often as any other."[34] When smaller samples from a larger class are drawn in this way, Peirce said that the smaller group would show the same characteristics of the larger group. He called this the "rule of induction."[35] Chapter 7 will delve into random assignment in the social sciences; today, we more typically think of randomly assigning people to different treatments, but, as Peirce's and others' early work showed, it is important to randomize as much as you can, including the stimuli in an experiment.

As an aside, to illustrate the scope of Peirce's abilities, he is also known for developing a theory of semiotics, signs and symbols—one of the classic theories still in use today in critical/cultural work.[36]

RONALD FISHER'S PLOTS AND TEA

Decades later, Ronald Aylmer Fisher revived and popularized the idea of random assignment of treatments, leading to its widespread use.[37] Fisher was an agricultural scientist outside London in the 1920s and '30s. He is credited with inventing the design of experiments, including many of the statistics, concepts, and procedures still in use today.[38] For example, Fisher developed **analysis of variance**, whose statistic, the **F test**, was named after him, and also the idea of **Latin Square**, a procedure for ordering treatments to

compensate for systematic error or control unintended variation instead of randomizing (more about this in chapter 7).[i] Fisher was also credited with proposing the probability values of .05 and .01 for statistical significance, which is a scientist's way of judging whether something deviates from chance.[39] A *p* value of .05 means there is a "one in twenty chance that the result is mistaken."[40]

Ronald Fisher

Wikimedia Commons

Fisher's job was to test various fertilizers on crops at the Rothamsted Agricultural Experimental Station in England. Because of this, many of the terms still used in experiments today have an agricultural basis, such as **split plot designs**, which actually comes from the plots in fields that were split into separate sections so that each could receive a different treatment. When Fisher arrived at the station, it had been standard to test a different fertilizer every year. But Fisher realized that each year brought inherent differences in rainfall, temperature, weed growth, drainage, and other factors.[41] He coined the term *confound* when he realized there was no way to separate the effects of the fertilizers from these other conditions that were out of his control.[42] To determine whether the outcome was due to the fertilizer or something else, he decided to include all of the treatments—in his case, the many different kinds of fertilizers—in the same experiment. By cutting up the fields into small plots, with the pieces divided up into different rows and each row given a different treatment, he reasoned that the yearly differences would apply to all the treatments, effectively "controlling" those confounding conditions.

In addition, Fisher also is credited with popularizing random assignment.[43] This also grew out of his efforts to control the various conditions when he realized that systematically assigning different fertilizers to the fields could not rule out confounding factors associated with the soil and fields themselves. To solve this, he randomly assigned the fields to receive different treatments. In social science today, we think of random assignment as applying to the subjects who are assigned to different treatments; for Fisher, the fields were

[i] As a geneticist, Fisher was also interested in evolution and eugenics—the idea that selective breeding could improve the human race. Eugenics had been discussed since early Greece and Rome, and was a respectable scientific topic in his day, supported by many including George Bernard Shaw and Alexander Graham Bell. Fisher helped found a eugenics society at Cambridge University. The idea was thoroughly discredited after being associated with the genocide policies of Nazi Germany. (Gregory Cochran and Henry Harpending, *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution* [New York: Basic Books, 2009]; Encyclopedia Britannica, Sir Ronald Aylmer Fisher, https://www.britannica.com/biography/Ronald-Aylmer-Fisher, accessed March 30, 2018; Famous Scientists, https://www.famousscientists.org/ronald-fisher/, accessed March 15, 2018.)

Vanderdecken, Wikimedia Commons

his subjects. In the sense of random assignment, "random" means by a chance procedure, such as flipping a coin, and ensures that each participant has an equal chance of being assigned to any of the treatment or control groups. This helps ensure that any systematic differences are equally distributed across the different groups, so that any differences can be attributed to the treatment, not something inherently different about the people in the groups. Chapter 7 will discuss random assignment and how it is achieved in more depth.

One of Fisher's more relatable experiments that used random assignment is the now-famous Lady Tasting Tea experiment. It also illustrates the idea of a **null hypothesis**, or the supposition that there will be no difference between those who get the treatment and those who do not. As the story goes, a woman who claimed she could tell if the milk was added before or after the tea was given four cups where the tea was added first, and four cups where the milk was added first, the order of which was randomized. The null hypothesis was that she could *not* tell them apart. The chance of someone guessing all eight correctly was one in seventy; the woman in the study purportedly got all eight correct.[44] The tea experiment was supposedly a summer afternoon of fun, not a scientific investigation with published results. Fisher described it in chapter 2 of his book *The Design of Experiments*.[45] Nowhere does he say if it was actually conducted or give the results, so much of this is folklore.[46] Experiments had been done for hundreds of years before Fisher, but they were idiosyncratic, varying with each experimenter. Fisher's was the first book that systematically codified how to do experiments.

B. F. SKINNER: SMALL SAMPLES, HIGH TECH

Fisher's principles were just becoming popular when B. F. (short for Burrhus Frederic, but friends called him Fred[47]) Skinner started studying the behavior of rats.[48] Skinner is popularly known for creating the Skinner box, a device he used to train rats to push a lever for food or to stop electric shocks that he used in developing his theory of operant conditioning.[ii] Not to belabor the details of his theory or studies here, the premise was

[ii]He developed another box for pigeons and one for babies; a student developed one for dogs, and much later, another student inspired by Skinner developed a human Skinner box—the study carrels of today. See Rutherford 2007, 2003.[56, 57]