4.4 Moving beyond simple experiments

Let's move beyond simple experiments. Three concepts are useful for rich experiments: validity, heterogeneity of treatment effects, and mechanisms.

Researchers who are new to experiments often focus on a very specific, narrow question: Does this treatment "work"? For example, does a phone call from a volunteer encourage someone to vote? Does changing a website button from blue to green increase the click-through rate? Unfortunately, loose phrasing about what "works" obscures the fact that narrowly focused experiments don't really tell you whether a treatment "works" in a general sense. Rather, narrowly focused experiments answer a much more specific question: What is the average effect of this specific treatment with this specific implementation for this population of participants at this time? I'll call experiments that focus on this narrow question *simple experiments*.

Simple experiments can provide valuable information, but they fail to answer many questions that are both important and interesting, such as whether there are some people for whom the treatment had a larger or smaller effect; whether there is another treatment that would be more effective; and whether this experiment relates to broader social theories.

In order to show the value of moving beyond simple experiments, let's consider an analog field experiment by P. Wesley Schultz and colleagues on the relationship between social norms and energy consumption (Schultz et al. 2007). Schultz and colleagues hung doorhangers on 300 households in San Marcos, California, and these doorhangers delivered different messages designed to encourage energy conservation. Then, Schultz and colleagues measured the effect of these messages on electricity consumption, both after one week and after three weeks; see figure 4.3 for a more detailed description of the experimental design.

The experiment had two conditions. In the first, households received general energy-saving tips (e.g., use fans instead of air conditioners) and information about their energy usage compared with the average energy usage in their neighborhood. Schultz and colleagues called this the *descriptive normative* condition because the information about the energy use in the neighborhood provided information about typical behavior (i.e., a descriptive norm). When Schultz and colleagues looked at the resulting energy usage

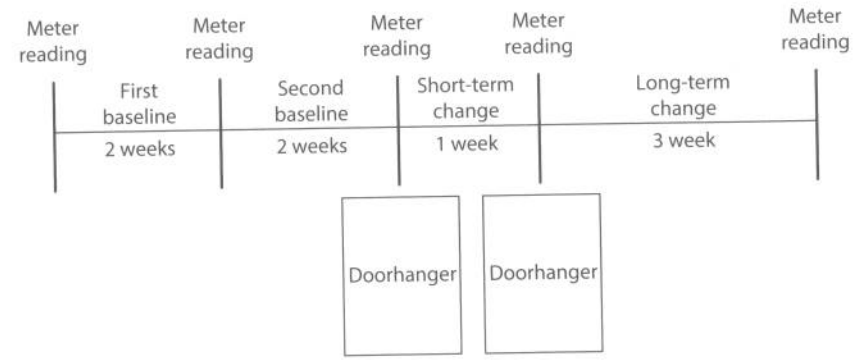


Figure 4.3: Schematic of the experimental design from Schultz et al. (2007). The field experiment involved visiting about 300 households in San Marcos, California five times over an eight-week period. On each visit, the researchers manually took a reading from the house's power meter. On two of the visits, they placed doorhangers on each house providing some information about the household's energy usage. The research question was how the content of these messages would impact energy use.

in this group, the treatment appeared to have no effect, in either the short or long term; in other words, the treatment didn't seem to "work" (figure 4.4).

Fortunately, Schultz and colleagues did not settle for this simplistic analysis. Before the experiment began, they reasoned that heavy users of electricity—people above the mean—might reduce their consumption, and that light users of electricity—people below the mean—might actually increase their consumption. When they looked at the data, that's exactly what they found (figure 4.4). Thus, what looked like a treatment that was having no effect was actually a treatment that had two offsetting effects. This counterproductive increase among the light users is an example of a *boomerang effect*, where a treatment can have the opposite effect from what was intended.

Simultaneous to the first condition, Schultz and colleagues also ran a second condition. The households in the second condition received the exact same treatment—general energy-saving tips and information about their household's energy usage compared with the average for their neighborhood—with one tiny addition: for people with below-average consumption, the researchers added a :) and for people with above-average consumption they added a :(. These emoticons were designed to trigger what the researchers called *injunctive norms*. Injunctive norms refer to perceptions of what is commonly approved (and disapproved), whereas descriptive

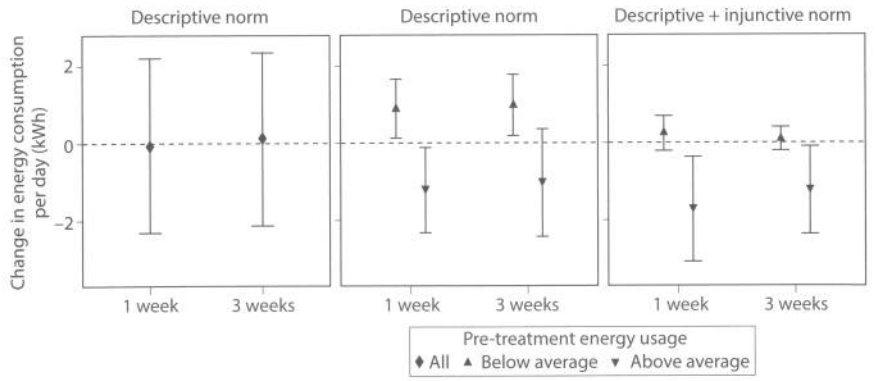


Figure 4.4: Results from Schultz et al. (2007). Panel (a) shows that the descriptive norm treatment has an estimated zero average treatment effect. However, panel (b) shows that this average treatment effect is actually composed of two offsetting effects. For heavy users, the treatment decreased usage, but for light users, the treatment increased usage. Finally, panel (c) shows that the second treatment, which used descriptive and injunctive norms, had roughly the same effect on heavy users but mitigated the boomerang effect on light users. Adapted from Schultz et al. (2007).

norms refer to perceptions of what is commonly done (Reno, Cialdini, and Kallgren 1993).

By adding this one tiny emoticon, the researchers dramatically reduced the boomerang effect (figure 4.4). Thus, by making this one simple change—a change that was motivated by an abstract social psychological theory (Cialdini, Kallgren, and Reno 1991)—the researchers were able to turn a program that didn't seem to work into one that worked, and, simultaneously, they were able to contribute to the general understanding of how social norms affect human behavior.

At this point, however, you might notice that something is a bit different about this experiment. In particular, the experiment of Schultz and colleagues doesn't really have a control group in the same way that randomized controlled experiments do. A comparison between this design and that of Restivo and van de Rijt illustrates the differences between two major experimental designs. In *between-subjects designs*, such as that of Restivo and van de Rijt, there is a treatment group and a control group. In *within-subjects designs*, on the other hand, the behavior of participants is compared before and after the treatment (Greenwald 1976; Charness, Gneezy, and Kuhn 2012). In a within-subjects experiment, it is as if each participant acts as her own control group. The strength of between-subjects designs is

that they provide protection against confounders (as I described earlier), while the strength of within-subjects experiments is increased precision of estimates. Finally, to foreshadow an idea that will come later when I offer advice about designing digital experiments, a *mixed design* combines the improved precision of within-subjects designs and the protection against confounding of between-subjects designs (figure 4.5).

Overall, the design and results of the study by Schultz and colleagues (2007) show the value of moving beyond simple experiments. Fortunately, you don't need to be a creative genius to design experiments like this. Social scientists have developed three concepts that will guide you toward richer experiments: (1) validity, (2) heterogeneity of treatment effects, and (3) mechanisms. That is, if you keep these three ideas in mind while you are designing your experiment, you will naturally create a more interesting and useful experiment. In order to illustrate these three concepts in action, I'll describe a number of follow-up partially digital field experiments that built on the elegant design and exciting results of Schultz and colleagues (2007). As you will see, through more careful design, implementation, analysis, and interpretation, you too can move beyond simple experiments.

4.4.1 Validity

> Validity refers to how much the results of an experiment support a more general conclusion.

No experiment is perfect, and researchers have developed an extensive vocabulary to describe possible problems. *Validity* refers to the extent to which the results of a particular experiment support some more general conclusion. Social scientists have found it helpful to split validity into four main types: statistical conclusion validity, internal validity, construct validity, and external validity (Shadish, Cook, and Campbell 2001, chapter 2). Mastering these concepts will provide you with a mental checklist for critiquing and improving the design and analysis of an experiment, and it will help you communicate with other researchers.

Statistical conclusion validity centers around whether the statistical analysis of the experiment was done correctly. In the context of Schultz et al. (2007), such a question might center on whether they computed their p-values correctly. The statistical principles need to design and analyze

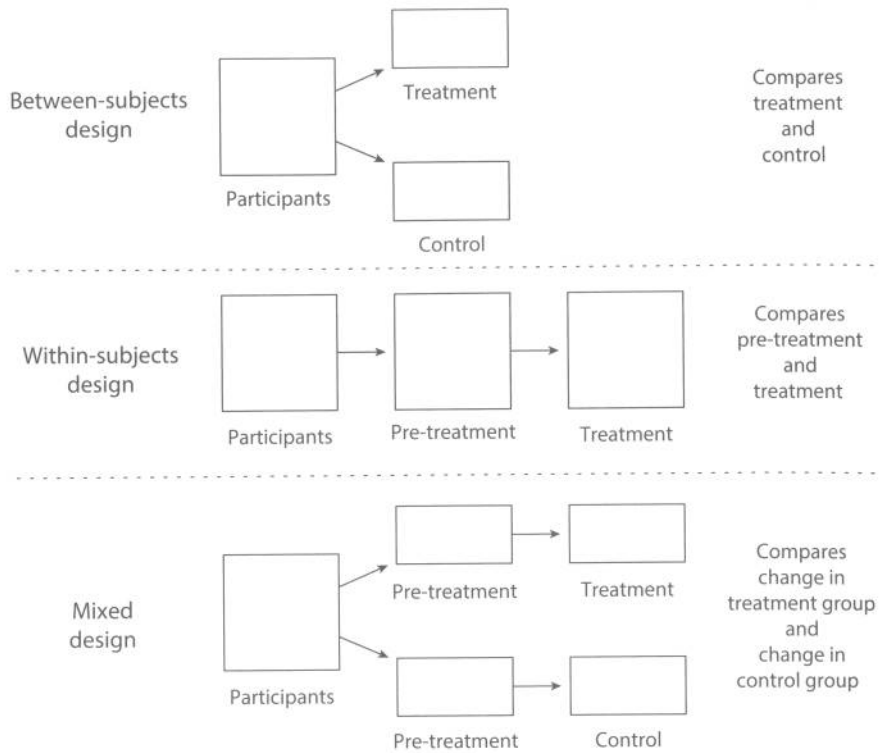


Figure 4.5: Three experimental designs. Standard randomized controlled experiments use *between-subjects* designs. An example of a between-subjects design is Restivo and van de Rijt's (2012) experiment on barnstars and contributions to Wikipedia: the researchers randomly divided participants into treatment and control groups, gave participants in the treatment group a barnstar, and compared outcomes for the two groups. The second type of design is a *within-subjects* design. The two experiments in Schultz and colleagues' (2007) study on social norms and energy use illustrate a within-subjects design: the researchers compared the electricity use of participants before and after receiving the treatment. Within-subjects designs offer improved statistical precision, but they are open to possible confounders (e.g., changes in weather between the pre-treatment and treatment periods) (Greenwald 1976; Charness, Gneezy, and Kuhn 2012). Within-subjects designs are also sometimes called repeated measures designs. Finally, *mixed designs* combine the improved precision of within-subjects designs and the protection against confounding of between-subjects designs. In a mixed design, a researcher compares the change in outcomes for people in the treatment and control groups. When researchers already have pre-treatment information, as is the case in many digital experiments, mixed designs are generally preferable to between-subjects designs because they result in improved precision of estimates.

experiments are beyond the scope of this book, but they have have not fundamentally changed in the digital age. What has changed, however, is that the data environment in digital experiments has created new opportunities such as using machine learning methods to estimate heterogeneity of treatment effects (Imai and Ratkovic 2013).

Internal validity centers around whether the experimental procedures were performed correctly. Returning to the experiment of Schultz et al. (2007), questions about internal validity could center around randomization, delivery of treatment, and measurement of outcomes. For example, you might be concerned that the research assistants did not read the electric meters reliably. In fact, Schultz and colleagues were worried about this problem, and they had a sample of meters read twice; fortunately, the results were essentially identical. In general, Schultz and colleagues' experiment appears to have high internal validity, but this is not always the case: complex field and online experiments often run into problems actually delivering the right treatment to the right people and measuring the outcomes for everyone. Fortunately, the digital age can help reduce concerns about internal validity, because it is now easier to ensure that the treatment is delivered to those who are supposed to receive it and to measure outcomes for all participants.

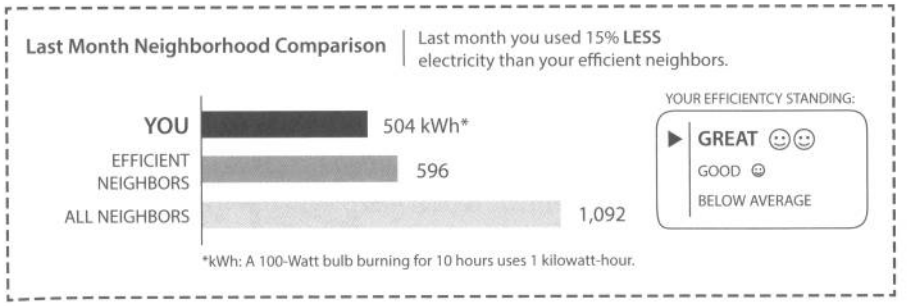
Construct validity centers around the match between the data and the theoretical constructs. As discussed in chapter 2, constructs are abstract concepts that social scientists reason about. Unfortunately, these abstract concepts don't always have clear definitions and measurements. Returning to Schultz et al. (2007), the claim that injunctive social norms can lower electricity use requires researchers to design a treatment that would manipulate "injunctive social norms" (e.g., an emoticon) and to measure "electricity use". In analog experiments, many researchers designed their own treatments and measured their own outcomes. This approach ensures that, as much as possible, the experiments match the abstract constructs being studied. In digital experiments where researchers partner with companies or governments to deliver treatments and use always-on data systems to measure outcomes, the match between the experiment and the theoretical constructs may be less tight. Thus, I expect that construct validity will tend to be a bigger concern in digital experiments than in analog experiments.

Finally, *external validity* centers around whether the results of this experiment can be generalized to other situations. Returning to Schultz et al. (2007), one could ask whether this same idea—providing people with

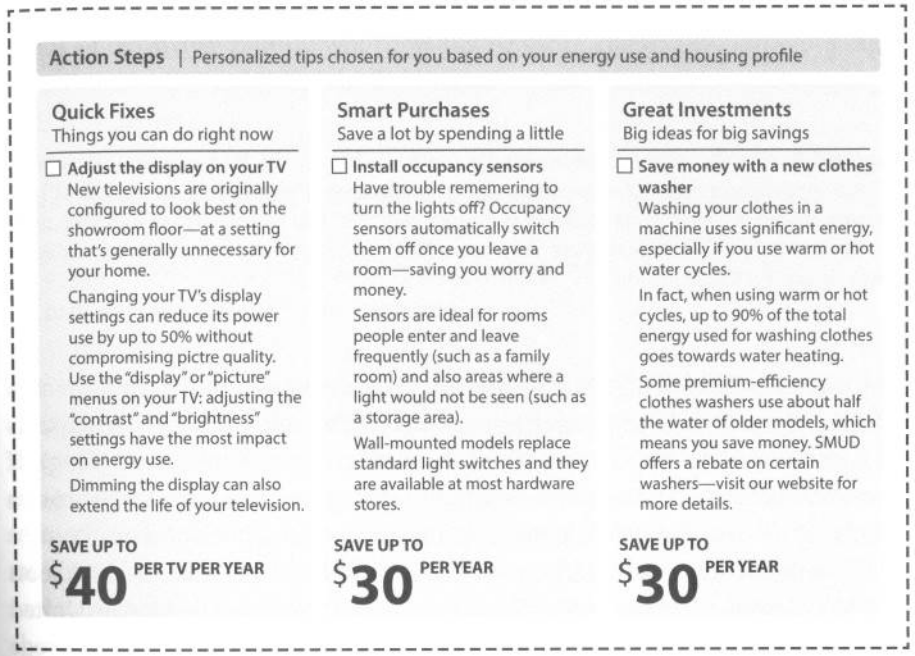
information about their energy usage in relationship to their peers and a signal of injunctive norms (e.g., an emoticon)—would reduce energy usage if it were done in a different way in a different setting. For most well-designed and well-run experiments, concerns about external validity are the hardest to address. In the past, these debates about external validity frequently involved nothing more than a group of people sitting in a room trying to imagine what would have happened if the procedures had been done in a different way, or in a different place, or with different participants. Fortunately, the digital age enables researchers to move beyond these data-free speculations and assess external validity empirically.

Because the results from Schultz et al. (2007) were so exciting, a company named Opower partnered with utilities in the United States to deploy the treatment more widely. Based on the design of Schultz et al. (2007), Opower created customized Home Energy Reports that had two main modules: one showing a household's electricity usage relative to its neighbors with an emoticon and one providing tips for lowering energy usage (figure 4.6). Then, in partnership with researchers, Opower ran randomized controlled experiments to assess the impact of these Home Energy Reports. Even though the treatments in these experiments were typically delivered physically—usually through old-fashioned snail mail—the outcome was measured using digital devices in the physical world (e.g., power meters). Further, rather than manually collecting this information by research assistants visiting each house, the Opower experiments were all done in partnership with power companies, enabling the researchers to access the power readings. Thus, these partially digital field experiments were run at a massive scale at low variable cost.

In a first set of experiments involving 600,000 households from 10 different sites, Allcott (2011) found that the Home Energy Report lowered electricity consumption. In other words, the results from the much larger, more geographically diverse study were qualitatively similar to the results from Schultz et al. (2007). Further, in subsequent research involving eight million additional households from 101 different sites, Allcott (2015) again found that the Home Energy Report consistently lowered electricity consumption. This much larger set of experiments also revealed an interesting new pattern that would not be visible in any single experiment: the size of the effect declined in the later experiments (figure 4.7). Allcott (2015) speculated that this decline happened because, over time, the treatment was being applied to



Peer Comparison Module



Action Steps Module

Figure 4.6: The Home Energy Reports had a Social Comparison Module and an Action Steps Module. Reproduced by permission from Elsevier from Allcott (2011), figures 1 and 2.

different types of participants. More specifically, utilities with more environmentally focused customers were more likely adopt the program earlier, and their customers were more responsive to the treatment. As utilities with less environmentally focused customers adopted the program, its effectiveness appeared to decline. Thus, just as randomization in experiments ensures that

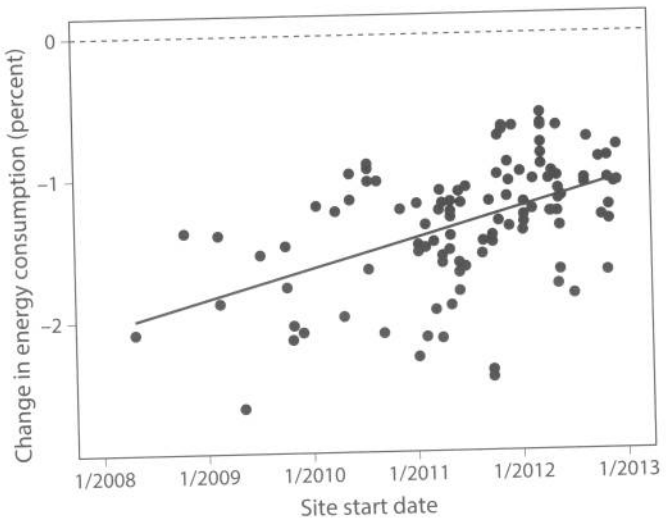


Figure 4.7: Results of 111 experiments testing the effect of the Home Energy Report on electricity consumption. At sites where the program was adopted later, it tended to have smaller effects. Allcott (2015) argues that a major source of this pattern is that sites with more environmentally focused customers were more likely to adopt the program earlier. Adapted from Allcott (2015), figure 3.

the treatment and control group are similar, randomization in research sites ensures that the estimates can be generalized from one group of participants to a more general population (think back to chapter 3 about sampling). If research sites are not sampled randomly, then generalization—even from a perfectly designed and conducted experiment—can be problematic.

Together, these 111 experiments—10 in Allcott (2011) and 101 in Allcott (2015)—involved about 8.5 million households from all over the United States. They consistently show that Home Energy Reports reduce average electricity consumption, a result that supports the original findings of Schultz and colleagues from 300 homes in California. Beyond just replicating these original results, the follow-up experiments also show that the size of the effect varies by location. This set of experiments also illustrates two more general points about partially digital field experiments. First, researchers will be able to empirically address concerns about external validity when the cost of running experiments is low, and this can occur if the outcome is already being measured by an always-on data system. Therefore, it suggests that researchers should be on the lookout for other interesting and important

behaviors that are already being recorded, and then design experiments on top of this existing measuring infrastructure. Second, this set of experiments reminds us that digital field experiments are not just online; increasingly, I expect that they will be everywhere, with many outcomes measured by sensors in the built environment.

The four types of validity—statistical conclusion validity, internal validity, construct validity, and external validity—provide a mental checklist to help researchers assess whether the results from a particular experiment support a more general conclusion. Compared with analog-age experiments, in digital-age experiments, it should be easier to address external validity empirically, and it should also be easier to ensure internal validity. On the other hand, issues of construct validity will probably be more challenging in digital-age experiments, especially digital field experiments that involve partnerships with companies.

4.4.2 Heterogeneity of treatment effects

Experiments normally measure the average effect, but the effect is probably not the same for everyone.

The second key idea for moving beyond simple experiments is *heterogeneity of treatment effects*. The experiment of Schultz et al. (2007) powerfully illustrates how the same treatment can have a different effect on different kinds of people (figure 4.4). In most analog experiments, however, researchers focused on average treatment effects because there were a small number of participants and little was known about them. In digital experiments, however, there are often many more participants and more is known about them. In this different data environment, researchers who continue to estimate only average treatment effects will miss out the ways in which estimates about the heterogeneity of treatment effects can provide clues about how a treatment works, how it can be improved, and how it can be targeted to those most likely to benefit.

Two examples of heterogeneity of treatment effects come from additional research on the Home Energy Reports. First, Allcott (2011) used the large sample size (600,000 households) to further split the sample and estimate the effect of the Home Energy Report by decile of pre-treatment energy usage. While Schultz et al. (2007) found differences between heavy and light users,

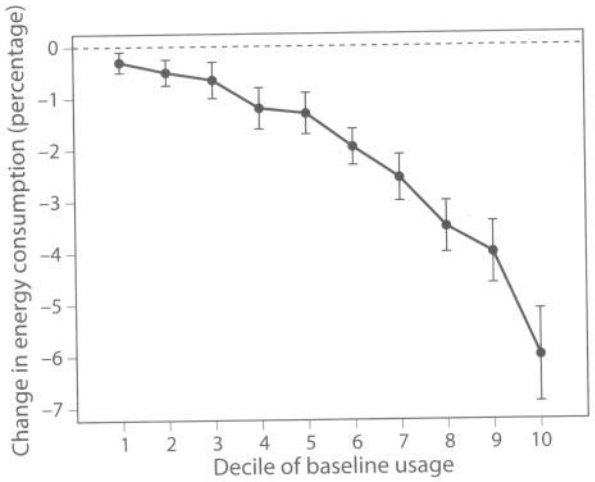


Figure 4.8: Heterogeneity of treatment effects in Allcott (2011). The decrease in energy use was different for people in different deciles of baseline usage. Adapted from Allcott (2011), figure 8.

Allcott (2011) found that there were also differences within the heavy- and light-user groups. For example, the heaviest users (those in the top decile) reduced their energy usage twice as much as someone in the middle of the heavy-user group (figure 4.8). Further, estimating the effect by pre-treatment behavior also revealed that there was no boomerang effect, even for the lightest users (figure 4.8).

In a related study, Costa and Kahn (2013) speculated that the effectiveness of the Home Energy Report could vary based on a participant's political ideology and that the treatment might actually cause people with certain ideologies to increase their electricity use. In other words, they speculated that the Home Energy Reports might be creating a boomerang effect for some types of people. To assess this possibility, Costa and Kahn merged the Opower data with data purchased from a third-party aggregator that included information such as political party registration, donations to environmental organizations, and household participation in renewable energy programs. With this merged dataset, Costa and Kahn found that the Home Energy Reports produced broadly similar effects for participants with different ideologies; there was no evidence that any group exhibited boomerang effects (figure 4.9).

As these two examples illustrate, in the digital age, we can move from esti-

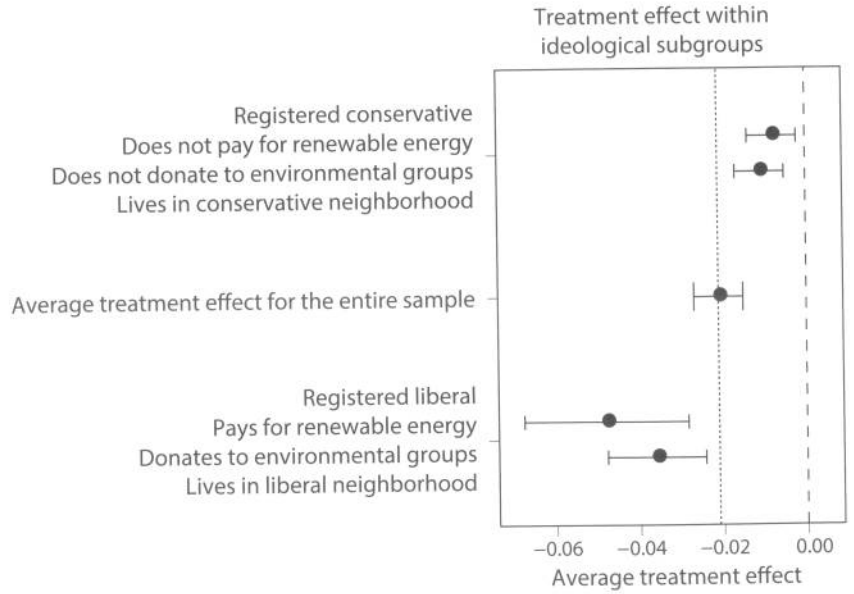


Figure 4.9: Heterogeneity of treatment effects in Costa and Kahn (2013). The estimated average treatment effect for the entire sample is −2.1% [−1.5%, −2.7%]. After combining information from the experiment with information about the households, Costa and Kahn (2013) used a series of statistical models to estimate the treatment effect for very specific groups of people. Two estimates are presented for each group because the estimates depend on the covariates they included in their statistical models (see models 4 and 6 in tables 3 and 4 in Costa and Kahn (2013)). As this example illustrates, treatment effects can be different for different people, and estimates of treatment effects that come from statistical models can depend on the details of those models (Grimmer, Messing, and Westwood 2014). Adapted from Costa and Kahn (2013), tables 3 and 4.

mating average treatment effects to estimating the heterogeneity of treatment effects, because we can have many more participants and we know more about those participants. Learning about heterogeneity of treatment effects can enable targeting of a treatment where it is most effective, provide facts that stimulate new theory development, and provide hints about possible mechanisms, the topic to which I now turn.

4.4.3 Mechanisms

Experiments measure what happened. Mechanisms explain why and how it happened.

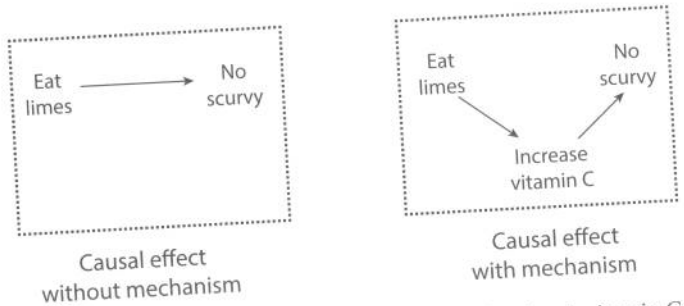


Figure 4.10: Limes prevent scurvy, and the mechanism is vitamin C.

The third key idea for moving beyond simple experiments is *mechanisms*. Mechanisms tell us *why* or *how* a treatment caused an effect. The process of searching for mechanisms is also sometimes called looking for *intervening variables* or *mediating variables*. Although experiments are good for estimating causal effects, they are often not designed to reveal mechanisms. Digital experiments can help us identify mechanisms in two ways: (1) they enable us to collect more process data and (2) they enable us to test many related treatments.

Because mechanisms are tricky to define formally (Hedström and Ylikoski 2010), I'm going to start with a simple example: limes and scurvy (Gerber and Green 2012). In the eighteenth century, doctors had a pretty good sense that when sailors ate limes, they did not get scurvy. Scurvy is a terrible disease, so this was powerful information. But these doctors did not know *why* limes prevented scurvy. It was not until 1932, almost 200 years later, that scientists could reliably show that vitamin C was the reason that lime prevented scurvy (Carpenter 1988, p. 191). In this case, vitamin C is the *mechanism* through which limes prevent scurvy (figure 4.10). Of course, identifying the mechanism is very important scientifically—lots of science is about understanding why things happen. Identifying mechanisms is also very important practically. Once we understand why a treatment works, we can potentially develop new treatments that work even better.

Unfortunately, isolating mechanisms is very difficult. Unlike limes and scurvy, in many social settings, treatments probably operate through many interrelated pathways. However, in the case of social norms and energy use, researchers have tried to isolate mechanisms by collecting process data and testing related treatments.

One way to test possible mechanisms is by collecting process data about how the treatment impacted possible mechanisms. For example, recall that Allcott (2011) showed that Home Energy Reports caused people to lower their electricity usage. But how did these reports lower electricity usage? What were the mechanisms? In a follow-up study, Allcott and Rogers (2014) partnered with a power company that, through a rebate program, had acquired information about which consumers upgraded their appliances to more energy-efficient models. Allcott and Rogers (2014) found that slightly more people receiving the Home Energy Reports upgraded their appliances. But this difference was so small that it could account for only 2% of the decrease in energy use in the treated households. In other words, appliance upgrades were not the dominant mechanism through which the Home Energy Report decreased electricity consumption.

A second way to study mechanisms is to run experiments with slightly different versions of the treatment. For example, in the experiment of Schultz et al. (2007) and all the subsequent Home Energy Report experiments, participants were provided with a treatment that had two main parts (1) tips about energy savings and (2) information about their energy use relative to their peers (figure 4.6). Thus, it is possible that the energy-saving tips were what caused the change, not the peer information. To assess the possibility that the tips alone might have been sufficient, Ferraro, Miranda, and Price (2011) partnered with a water company near Atlanta, Georgia, and ran a related experiment on water conservation involving about 100,000 households. There were four conditions:

- a group that received tips on saving water
- a group that received tips on saving water plus a moral appeal to save water
- a group that received tips on saving water plus a moral appeal to save water plus information about their water use relative to their peers
- a control group

The researchers found that the tips-only treatment had no effect on water usage in the short (one year), medium (two years), and long (three years) term. The tips plus appeal treatment caused participants to decrease water usage, but only in the short term. Finally, the tips plus appeal plus peer information treatment caused decreased usage in the short, medium,

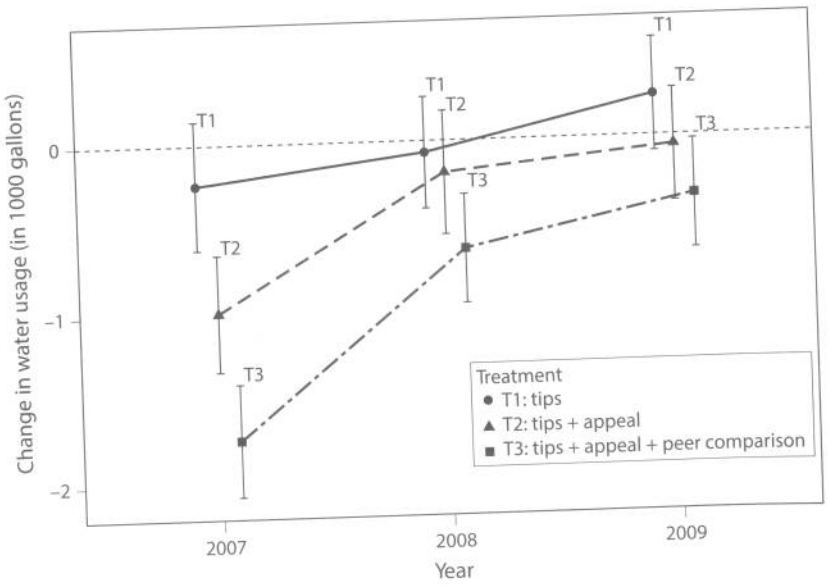


Figure 4.11: Results from Ferraro, Miranda, and Price (2011). Treatments were sent May 21, 2007, and effects were measured during the summers of 2007, 2008, and 2009. By unbundling the treatment, the researchers hoped to develop a better sense of the mechanisms. The tips-only treatment had essentially no effect in the short (one year), medium (two years), and long (three years) term. The tips plus appeal treatment caused participants to decrease water usage, but only in the short term. The advice plus appeal plus peer information treatment caused participants to decrease water usage in the short, medium, and long term. Vertical bars are estimated confidence intervals. See Bernedo, Ferraro, and Price (2014) for actual study materials. Adapted from Ferraro, Miranda, and Price (2011), table 1.

and long term (figure 4.11). These kinds of experiments with unbundled treatments are a good way to figure out which part of the treatment—or which parts together—are the ones that are causing the effect (Gerber and Green 2012, section 10.6). For example, the experiment of Ferraro and colleagues shows us that water-saving tips alone are not enough to decrease water usage.

Ideally, one would move beyond the layering of components (tips; tips plus appeal; tips plus appeal plus peer information) to a full factorial design—also sometimes called a 2^k factorial design—where each possible combination of the three elements is tested (table 4.1). By testing every possible combination of components, researchers can fully assess the effect of each component in isolation and in combination. For example, the experiment

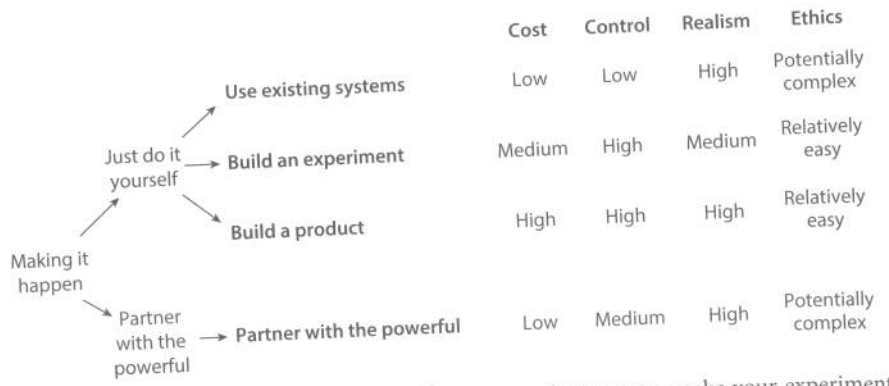
Table 4.1: Example of Treatments in a Full Factorial Design with Three Elements: Tips, Appeal, and Peer Information

Treatment	Characteristics
1	Control
2	Tips
3	Appeal
4	Peer information
5	Tips + appeal
6	Tips + peer information
7	Appeal + peer information
8	Tips + appeal + peer information

of Ferraro and colleagues does not reveal whether peer comparison alone would have been sufficient to lead to long-term changes in behavior. In the past, these full factorial designs have been difficult to run because they require a large number of participants and they require researchers to be able to precisely control and deliver a large number of treatments. But, in some situations, the digital age removes these logistical constraints.

In summary, mechanisms—the pathways through which a treatment has an effect—are incredibly important. Digital-age experiments can help researchers learn about mechanisms by (1) collecting process data and (2) enabling full factorial designs. The mechanisms suggested by these approaches can then be tested directly by experiments specifically designed to test mechanisms (Ludwig, Kling, and Mullainathan 2011; Imai, Tingley, and Yamamoto 2013; Pirlott and MacKinnon 2016).

In total, these three concepts—validity, heterogeneity of treatment effects, and mechanisms—provide a powerful set of ideas for designing and interpreting experiments. These concepts help researchers move beyond simple experiments about what "works" to richer experiments that have tighter links to theory, that reveal where and why treatments work, and that might even help researchers design more effective treatments. Given this conceptual background about experiments, I'll now turn to how you can actually make your experiments happen.



	Cost	Control	Realism	Ethics
Use existing systems	Low	Low	High	Potentially complex
Build an experiment	Medium	High	Medium	Relatively easy
Build a product	High	High	High	Relatively easy
Partner with the powerful	Low	Medium	High	Potentially complex

Figure 4.12: Summary of trade-offs for different ways that you can make your experiment happen. By *cost* I mean cost to the researcher in terms of time and money. By *control* I mean the ability to do what you want in terms of recruiting participants, randomization, delivering treatments, and measuring outcomes. By *realism* I mean the extent to which the decision environment matches those encountered in everyday life; note that high realism is not always important for testing theories (Falk and Heckman 2009). By *ethics* I mean the ability of well-intentioned researchers to manage ethical challenges that might arise.

4.5 Making it happen

Even if you don't work at a big tech company you can run digital experiments. You can either do it yourself or partner with someone who can help you (and who you can help).

By this point, I hope that you are excited about the possibilities of doing your own digital experiments. If you work at a big tech company, you might already be doing these experiments all the time. But if you don't work at a tech company, you might think that you can't run digital experiments. Fortunately, that's wrong: with a little creativity and hard work, everyone can run a digital experiment.

As a first step, it is helpful to distinguish between two main approaches: doing it yourself or partnering with the powerful. And there are even a few different ways that you can do it yourself: you can experiment in existing environments, build your own experiment, or build your own product for repeated experimentation. As you'll see from the examples below, none of these approaches is best in all situations, and it's best to think of them as offering trade-offs along four main dimensions: cost, control, realism, and ethics (figure 4.12).

4.5.1 Use existing environments

You can run experiments inside existing environments, often without any coding or partnership.

Logistically, the easiest way to do a digital experiment is to overlay your experiment on top of an existing environment. Such experiments can be run at a reasonably large scale and don't require partnership with a company or extensive software development.

For example, Jennifer Doleac and Luke Stein (2013) took advantage of an online marketplace similar to Craigslist in order to run an experiment that measured racial discrimination. They advertised thousands of iPods, and by systematically varying the characteristics of the seller, they were able to study the effect of race on economic transactions. Further, they used the scale of their experiment to estimate when the effect was bigger (heterogeneity of treatment effects) and to offer some ideas about why the effect might occur (mechanisms).

Doleac and Stein's iPod advertisements varied along three main dimensions. First, the researchers varied the characteristics of the seller, which was signaled by the hand photographed holding the iPod [white, black, white with tattoo] (figure 4.13). Second, they varied the asking price [$90, $110, $130]. Third, they varied the quality of the ad text [high-quality and low-quality (e.g., cApitalization errors and spelin errors)]. Thus, the authors had a $3 \times 3 \times 2$ design, which was deployed across more than 300 local markets, ranging from towns (e.g., Kokomo, Indiana and North Platte, Nebraska) to mega-cities (e.g., New York and Los Angeles).

Averaged across all conditions, the outcomes were better for the white sellers than the black sellers, with the tattooed sellers having intermediate results. For example, the white sellers received more offers and had higher final sale prices. Beyond these average effects, Doleac and Stein estimated the heterogeneity of effects. For example, one prediction from earlier theory is that discrimination would be less in markets where there is more competition between buyers. Using the number of offers in that market as a measure of the amount of buyer competition, the researchers found that black sellers did indeed receive worse offers in markets with a low degree of competition. Further, by comparing outcomes for the ads with high-quality and low-quality text, Doleac and Stein found that ad quality did not impact the