Citation in APA format: Wilson, T. D., Westgate, E. C., & Aronson, E. (2025). The art of experimentation in the laboratory and online. In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The handbook of social psychology* (6th ed.). Situational Press. https://doi.org/10.70400/UUQK3116

The Art of Experimentation in the Laboratory and Online

Timothy D. Wilson, University of Virginia

Erin C. Westgate, University of Florida

Elliot Aronson, University of California, Santa Cruz

If you are reading this chapter, you may well be new to the field of social psychology, perhaps at the beginning of your graduate career. If so, congratulations and welcome! We are confident that there are exciting times ahead for you as you develop your ideas and formulate novel hypotheses about social behavior. As we're sure you know, however, making predictions about how people will behave is merely the first step. Transforming these ideas into hypotheses that can be tested with sound scientific methods is the real challenge. As with any challenge, it can be both frustrating and great fun. As Leon Festinger once said, it is like solving a difficult puzzle:

> I love games. I think I could be very happy being a chess player or dealing with some other kinds of games. But I grew up in the Depression. It didn't seem one could survive on chess, and science is also a game. You have very strict ground rules in science, and your ideas have to check out with the empirical world. That's very tough and also very fascinating (quoted in Cohen, 1977, p. 133).

In this chapter we hope to convey some of this fascination with a particular kind of scientific game, namely the experimental method. This is not the only method available to social psychologists, of course. Researchers make good use of correlational designs and qualitative methods to explore social phenomena. And, as new sources of data have become available to social psychologists—including social media posts, financial transactions, video from surveillance cameras, and personal data from cell phones—there has been a corresponding increase in the methods available to analyze these data, including text analysis software and machine learning programs. We are excited by these developments and are pleased to see new chapters in this Handbook that discuss them.

The old should not be thrown out with the new, however, and here we advocate for the venerable experimental method, in which participants are randomly assigned to receive some form of "treatment," followed by a measurement of the effects of that treatment. To be sure, a great deal can be learned from other methods, but if the goal is to test causal hypotheses--that some psychological variable X is influencing an outcome Y--the experiment is the method of choice.

We have two main goals in this chapter. First, we will discuss why the experiment is the gold standard in psychology. This section may seem repetitive to experienced readers who already learned about the value of the experimental method, and indeed may already be conducting experiments of their own. It is worth revisiting the advantages and disadvantages of experiments, however, because the use of laboratory experiments has become less frequent in many areas, including social psychology (Ellsworth, 2010; Ross, in press). One reason for this is that social psychologists have ventured into areas in which it is more difficult to do experiments, such as the study of culture, close relationships, the areas of the brain that correspond to social cognition and behavior (social neuroscience), and investigations of "big data" gleaned from the internet and other sources. Another reason is that sophisticated statistical techniques (e.g., causal machine learning) are now available, allowing more precise tests of the relationships between variables in correlational designs. A third reason is a relatively new development, namely the ability to conduct experiments on the internet. Indeed, versions of this chapter that appeared in previous editions of *The Handbook of Social Psychology* focused on laboratory experiments, which involve "live" interactions with participants (e.g., Aronson, Wilson, & Brewer, 1998; Wilson, Aronson, & Carlsmith, 2010). We have expanded our discussion to include online experiments with remote participants, given how common such studies have become.

The second part of the chapter is more of a "how-to" manual describing in some detail how to conduct an experiment. We hasten to add that the best way to learn to do an experiment is to do so under the guidance of an expert, experienced researcher. Experimentation is very much a trade, like plumbing or carpentry or directing a play; the best way to learn to do it is by working alongside an expert in the craft. Nonetheless, just as it helps to read manuals explaining how to fix a leaky faucet or stage a production of Hamlet, our "how to do an experiment" manual might prove to be a helpful adjunct to a hands-on apprenticeship.

WHY DO LABORATORY EXPERIMENTS?

Varied sources of data are increasingly available, as are methods to analyze them. Why choose one over another? The obvious answer is that our methods should be dictated by the questions we seek to answer--and in science, those questions are typically about description, prediction, or causality. Sometimes scientists want to get the "lay of the land," namely to describe the nature of a phenomenon, in which case the observational method is appropriate. Scientists observe the phenomenon in as unbiased a way as possible, to understand as much about it as they can. In psychology, this can involve qualitative methods, whereby researchers attempt to understand people's experiences from their point of view, typically with in-depth interviews. One study, for example, conducted interviews with 22 British Muslim women, and yielded insights about how wearing hijab related to their sometimes conflicting religious, gender, and national identities (Hopkins & Greenwood, 2013).

Other times scientists want to be able to predict the occurrence of a phenomenon, by measuring two or more variables and computing the correlation(s) between them. Political scientists try to predict the outcome of an election from polling data, and economists try to predict the likelihood of a recession from the yield rates of Treasury bonds. Examples in psychology include predicting behavior from measures of personality traits, or from people's natural use of language via text analysis programs (Charlesworth et al., 2021; Eichstaedt et al., 2021; Pennebaker, 2011). Sophisticated statistical tools such as machine learning techniques can sort out relationships between multiple variables in large datasets, although their usefulness for addressing social psychological questions is the subject of debate (see Kosinski, 2025). For our purposes, the point is that the goal of machine learning is prediction, e.g., predicting from large datasets such varied behaviors as job performance, binge drinking, and the likelihood of donating to charities (Jacobucci & Grimm, 2020; Yarkoni & Westfall, 2017).

As important as prediction is, it has not been the primary goal of social psychologists. Rather, the goal is causal inference--*why* do people think, feel, and behave the way they do? As Yarkoni and Westfall (2017) put it, "Psychology has historically been concerned, first and foremost, with explaining the causal mechanisms that give rise to behavior" (p. 1100). Social psychologists address such classic causal questions as the mechanisms that lead to prejudice and stereotyping, susceptibility to social influence, the ways in which people process information about themselves and the social environment, and so on. And to do so, the experiment has been—and should continue to be—the method of choice.

To understand why, consider a classic social psychology experiment that has been replicated numerous times: The induced compliance dissonance experiment. This experiment, originally performed by Festinger and Carlsmith (1959), was designed to test what at the time was a provocative new formulation, cognitive dissonance theory. The theory argued that when people act contrary to their beliefs, for no apparent reason, psychological tension is produced. People can reduce this tension by convincing themselves that their behavior actually wasn't inconsistent with their beliefs, but instead reflected how they really felt.

To test this hypothesis, Festinger and Carlsmith (1959) asked participants to engage in two exceedingly boring tasks, each for 30 minutes: First participants were instructed to put 12 spools on a tray, remove the spools, and then fill up the tray again--time after time. Next participants repeatedly turned wooden pegs a quarter turn clockwise. When they were done, the experimenter explained that there were two conditions in the study. In one, which the participant had just completed, people did the tasks with no advance knowledge of what they entailed. In the other, an accomplice of the experimenter played the role of a prior participant and told the next participant that the task was really quite interesting--to see whether this affected their enjoyment of the task. Then, looking flustered, the experimenter mentioned that his accomplice had just phoned to say that they couldn't make it that day. Might the participant be willing to fill in and tell the next person that the task was enjoyable and fun? The experimenter was persuasive, and virtually all participants agreed to this request. As instructed, they went to the waiting room and told the next participant (actually an accomplice) that filling a tray with spools and turning wooden pegs was a lot of fun. So, the stage was set: Participants had acted contrary to their beliefs--telling someone that the boring task was quite interesting. The question was, would they come to believe the lie they had just told?

The researchers hypothesized that dissonance would only occur if people could not find a good reason for lying. So, in one condition, they provided that reason: They paid participants $20 to tell the lie and for "being on call" for future sessions (which was quite a hefty sum in 1959—equivalent to about $200 today). In that case, they reasoned, people would not feel any dissonance about lying, because they had a perfectly good reason for doing so: they'd been paid a substantial sum and were helping the experimenter. In another condition, participants were paid only $.50 to tell the lie. These participants should experience dissonance, the researchers reasoned, because they lied for no good reason (the pay was trivial). How might this dissonance be reduced? By convincing themselves that they hadn't really lied because the task actually *was* kind of interesting. And that's exactly what happened: People who lied for $20 reported that the task was unenjoyable, just like participants in a control group who did the task but were not asked to lie. People who lied for $.50, however, reported that the task was significantly *more* enjoyable than participants in the other two groups. How could this be? Presumably, as a way of reducing their cognitive dissonance, these participants came to believe their own lies.

Some readers might find this study to be flawed in several ways. It used college students as participants, who are not representative of the population at large. It was done in the artificial confines of the laboratory, where participants knew they were in a psychology study. It involved deception: the cover story about needing someone to lie to the next participant was a ruse, and an accomplice of the experimenter played the role of the next participant. Why go to all this trouble? Was it even ethical?

The contemporary student of social psychology might have a different set of misgivings. The study seems very inefficient; participants were run one at a time in the laboratory, as opposed, say, to collecting data from hundreds of online participants in a matter of hours. The study seems underpowered; there were only 20 participants in each of the three conditions. And, it *was* a lot of trouble--an elaborate cover story, an accomplice playing the role of the next participant, the use of deception, finding a supply of spools and wooden pegs. The whole study may well seem quaint and old-fashioned. Wouldn't it be a lot easier to run the study online with mTurk participants, and get the data in a day or two?

These are all good questions that merit thoughtful responses. To begin, we acknowledge that the Festinger and Carlsmith study was elaborate and difficult to run, which was recognized at the time it was published. Robert Cohen at Yale University invented a way of testing the same hypotheses with a much simpler procedure, which has become known as the "counterattitudinal essay" experiment (reported in Brehm & Cohen, 1962). Students at Yale University were asked to write an essay in support of a position they disagreed with, namely that the local police were justified in using brutal tactics at a recent student protest. Students uniformly felt that the police had used excessive force; thus, asking them to write an essay supporting the police was contrary to their attitudes. Cohen varied how much participants were paid to write the essay: $.50, $1.00, $5.00, or $10.00. Then, when they were done, participants reported how justified they felt the police actions had been, and the researchers compared their answers to those of students in a control group who did not write an essay. Consistent with cognitive dissonance theory, the less participants were paid, the *more* they changed their attitude in favor of the police. Writing the essay for no good reason (low external justification) presumably induced cognitive dissonance, which participants reduced by changing their attitudes to be consistent with what they had written. Like the students in Festinger & Carlsmith (1959), they came to believe their own lies.

The counterattitudinal essay experiment is easier to conduct than the elaborately staged Festinger and Carlsmith study, which probably explains why it has been repeated and extended dozens of times. But the question remains of why researchers make the effort to conduct such laboratory studies, especially nowadays when there is a ready availability of online participants.

Let's consider other ways the researchers could have tested their hypothesis about cognitive dissonance, or how we might test it today. One simple approach would be to find a setting in which people are publicly voicing their opinion about an important issue, and then assessing whether their private beliefs match their public statements. We could attend a local school board meeting, for example, and listen to public comments on a highly politicized topic, such as whether schools should be allowed to teach kids about systemic racism in America or whether they should remove such material from the curriculum. We could note what position each speaker took, and then survey them, after the meeting, on how much they supported (or did not support) the curriculum change. If the "saying is believing" hypothesis about cognitive dissonance is correct, then people's privately reported attitudes on the survey should align closely with the position they voiced publicly.

This study would have several advantages over the ones conducted by the early dissonance theorists. It would be done in the real world, and assess people's attitudes toward a current topic of considerable importance. We hope you can see, however, the limitations of such a study. Dissonance theorists were testing a causal hypothesis, namely that people who freely voice an opinion contrary to their beliefs, without good justification, come to believe what they were saying. In our hypothetical school board study, we have no way of knowing whether people's public comments about the curriculum changed their beliefs, or whether they were simply expressing what they already believed to begin with. More than likely, people were voicing opinions that were entirely consistent with how they already felt, meaning that no attitude change was occurring. People opposed to teaching about systemic racism likely voiced that opinion publicly at the meeting *and* reported that attitude privately on our survey.

In other words, our hypothetical school board study doesn't capture the psychological conditions necessary to test the dissonance hypothesis. OK, we might respond, let's try a different way. Like Cohen in his study at Yale, let's select a sample of people who we already know feel a certain way about an issue. We can be pretty sure, for example, that college students are opposed to major tuition increases at their university. In an online study, we could tell students that a tuition increase is under consideration, and then ask them to help us out by writing an essay in support of the increase. Does writing the essay lead to more support for the increase, as dissonance theory predicts? Let's say that it does: About half the people agree to write the essay, and as predicted, they subsequently voice more support for the increase than those who declined to write the essay.

Correlational Versus Experimental Studies

To some, this hypothetical study will seem superior to the ones conducted by early dissonance theorists: It is much simpler, can be conducted quickly online, and is void of deception. But as the careful reader will have noted, it employed a correlational design. The study measured whether participants wrote the essay and their subsequent attitude toward a tuition increase, and found that the former was correlated with the latter. But as any student of science knows, correlation does not equal causation. Writing the essay may not have caused support for the tuition increase at all; indeed, the causal direction might have been the opposite: Those who were most favorable toward a tuition increase may have been more likely to agree to write the essay. Unmeasured third variables (e.g., a favorable attitude toward the university) might have produced more compliance with the researchers' request to write the essay and correlate with more favorable attitudes toward a tuition increase.

That is, like most social psychological researchers, the early dissonance theorists were testing a causal hypothesis, namely that one psychological variable (acting contrary to one's attitudes) would have an interesting effect on another (attitude change), under specified conditions (e.g., when external justification for the behavior was low).

But when there are different experimental conditions, how do we know that any differences found are due to the experimental manipulation, and not to pre-existing differences in the participants? In the Festinger and Carlsmith (1959) study, for example, participants paid $.50 to lie reported that the task was more enjoyable than did participants in the control condition, who did not lie. How do we know that this was because of lying for a small amount of money? The participants in the lie-for-$.50 condition may have differed from the participants in the control condition in numerous ways that accounted for the differences, e.g., maybe those in the former condition had taken high school shop classes in which they crafted wooden spools and pegs, which made them admire the handiwork of the researchers and enjoy the task more.

Festinger and Carlsmith (1959) solved this problem with the most important advantage of experimental designs: The ability to randomly assign people to conditions. They randomly assigned people to lie for $.50, to lie for $20, or to not lie at all. By so doing, they could be reasonably sure that there were no differences between the conditions in participants' backgrounds, personalities, experiences in high school shop, or whatever. Random assignment is the great equalizer: As long as the sample size is sufficiently large, researchers can be relatively certain that all characteristics of participants are distributed evenly across conditions. Any differences that are observed, then, are likely to be due to the independent variable encountered in the experiment. For instance, rather than letting participants choose whether (or not) to write an essay in support of tuition increases, a better research design would have been to randomly assign some students to write an essay in support of the increase, while others wrote an essay in opposition to the increase (or no essay at all). In this way, we could be sure that there were no pre-existing differences between the types of students who wrote the essay and those who did not.

Our discussion of the limits of correlational designs--and the advantage of experiments--is no different from that in any introductory course in statistics or research methodology. Indeed, we trust that most readers of this chapter are familiar with the edict that correlation does not equal causation—in the abstract. In practice, however, even seasoned researchers sometimes overlook this truism. We often see students propose correlational studies to test causal hypotheses--even advanced graduate students. As journal reviewers, we see authors submit a paper claiming a causal link between two variables (e.g., between income and happiness), when they report only correlational data.

Consider, for example, the following two (fictitious) investigations of the same problem. In the first, a team of researchers finds that school performance in a large sample of children from a low-SES background is related to the frequency with which they eat breakfast in the morning. The more often the kids eat breakfast, the better their school performance, with a highly significant correlation of .30 (this means that the relationship between eating breakfast and school performance is moderately strong and highly unlikely to have occurred by chance). As far as you can tell, the researchers used good measures and the study was well conducted. What do you think of this finding? Does it make you more confident that programs that provide free breakfasts for underprivileged children are having positive effects on their academic performance? If you were reviewing a report of this study for a journal, how likely would you be to recommend publication? Most of us, we suspect, would find this to be an interesting and well-conducted study that should be in the literature.

Now consider this study: A team of researchers conducts an experiment with a large sample of children from low-SES backgrounds. Half of the kids are randomly assigned to a condition in which they receive free breakfasts at school every morning, whereas the other half are in a control group that does not receive this intervention. Unfortunately, the researchers introduced a confound into their design: While the kids in the first group eat their breakfast, teachers also read to them and help them with their homework. After a few months, the researchers assess the kids' school performance and find that those in the breakfast condition are doing significantly better than those in the control condition. The measure of academic performance is the same as in the previous study and the magnitude of the effect is the same. What do you think of this experiment? How likely would you be to recommend that it be published? The confound in the design, we would guess, is likely to be glaring and appalling to most of us. Is it eating breakfast that improved the kids' performance or the reading and extra attention from the teachers? Many of us would feel that the design of this study is so flawed that it should not be published.

But let's compare the two studies more carefully. The key question is how confident we can be that eating breakfast causes improved academic performance. The flaw in the experiment is that we cannot be sure whether eating breakfast or extra attention from a teacher or both were responsible for the improved performance. But how confident can we be from the correlational study? Kids who eat breakfast probably differ in countless ways from kids who do not. They may come from different families, get more sleep--or, for that matter, have parents or teachers who are more likely to help them with their homework! The experimental study, despite its flaw, rules out every single one of these alternative explanations except for one. Admittedly, this is a serious flaw; the researchers did err by confounding breakfast eating with extra attention from the teachers. But the fact remains that the correlational study leaves the door open to the same confound, and dozens or hundreds of others besides. If the goal is to reduce uncertainty about causality, surely the correlational study is much more flawed than the experimental one. Why, then, does it seem like more can be learned from the correlational study? One reason is that the correlational study was done well, by the standards of correlational designs, whereas the experimental study was done poorly, by the standards of experimental designs. Our point is that the same standard should be applied to both types of studies: How much do they reduce uncertainty about causality?

The ability to determine relationships between variables in correlational designs has improved, we should add, with the advent of sophisticated statistical techniques such as RI-CLPM, causal machine learning, and network psychometrics, among many others (e.g., Epskamp et al., 2018; Orth et al., 2021; Schölkopf et al., 2021). These methods offer useful techniques for testing complex relationships between several variables. However, they cannot, in the absence of experimental manipulations with random assignment, determine causal relationships. To understand why, we must look at the three criteria for causality: first, our two variables must be correlated or co-occur. Consider a summer storm: Physics tells us that lightning causes thunder when it heats the air and causes it to expand. As such, lightning and thunder usually occur in close proximity. Second, the causal variable X needs to occur before the outcome variable Y. This is the principle of temporal precedence; lightning can only cause thunder if the lightning came first. If we sometimes hear thunder before we see a lightning bolt, we naturally reason that the lightning bolt can't be causing that particular roll of thunder. Finally, we must be able to rule out any other competing explanations (exclusivity). Could something else (like a rumbling truck) be causing those loud sounds--or is lightning the only possibility? While correlational studies can do an excellent job of meeting criterion #1 (co-occurrence) and, in longitudinal designs, criterion #2 (temporal precedence), they cannot meet criterion #3 (exclusivity), even when advanced statistical models are employed. One reason for this is obvious but sometimes overlooked: It is impossible to measure all variables in a correlational design, and the researchers might have omitted one or more crucial causal variables. Thus, although a given model may suggest a direct relationship between two variables, one can never be sure whether this is because one variable really causes the other or whether there are unmeasured variables that are the true causes and happen to correlate highly with the measured variables. To be sure, researchers can think carefully about (and attempt to measure) likely confounds. Tools such as Directed Acyclic Graphs (DAGs) and Theory Maps help researchers think through which third variables might plausibly influence both predictor and outcome and should be measured and controlled for (e.g., Gray, 2017; Rohrer, 2018). Such tools may also guide the selection of matched controls in pseudo-experimental designs or instrumental variables that can be used to generate estimates of causal effects. And recent evidence shows that such causal inference models can perform quite well (e.g., Concato et al., 2000; Ferraro & Miranda, 2016; Franklin et al., 2021; Wu et al., 2022). However, such tools are only as good as our own causal reasoning--and as we know, lay theories of causality often get it wrong, holding that some things influence outcomes (that don't) or that other factors are irrelevant (that aren't). In short, no statistical technique can control for variables that weren't measured, and it is impossible to measure everything (e.g., Gordon et al., 2019). The only way to definitively rule out such alternative explanations is to use experimental designs, in which people are randomly assigned to different experimental conditions.

Finally, we note that even in those circumstances in which experimental designs might be impractical or unethical (e.g., understanding the causal effects of parenthood on happiness), controlling statistically for potential confounds is more difficult than it seems. Perfect measurement is impossible for most constructs of interest to psychologists--people's scores on self-report questionnaires are influenced by any number of factors beyond those we want to measure, and are imperfect (if useful!) representations of their underlying constructs. This measurement error, when not properly accounted for by techniques such as Structural Equation Modeling (SEM), can lead to serious errors in causal inference. For example, simply adding a potential third variable as a covariate to a multiple regression, as is commonly done, does not rule out the possibility that variance in the underlying construct may "leak" out and affect the results. That is, even when researchers correctly deduce that a given variable might be acting as a confound, and measure and control for it in their models, that confound may still continue to artificially inflate the observed relationship between predictor and outcome, if measured with error (see Westfall & Yarkoni, 2016 for an excellent discussion of this issue and potential solutions).

Failures to appreciate the limits of correlational designs have had tragic consequences. In a well-known example, nurses who took estrogen to alleviate the unpleasant side effects of menopause (e.g., hot flashes) were found to have fewer heart attacks. Based on this and other correlational studies, thousands of women undergoing menopause were prescribed estrogen as part of hormone replacement therapy. Only later, when researchers conducted a randomized controlled trial (RCT), randomly assigning some women to receive treatment and others to an untreated control group, was it discovered that estrogen actually increased the women's risk of heart disease (Rossouw, 2014). How can this be? In the correlational study, women from higher socioeconomic backgrounds were both more likely to seek out hormone replacement therapy and to receive better medical care and generally carry fewer risk factors for heart disease. And while some earlier studies tried to control for socioeconomic status, measurement error meant that they were unable to do so sufficiently in correlational designs.

Validity And Realism In Experiments

We hope we have convinced the reader of the great advantage of the experiment--its ability to answer causal questions. Some, however, might still be a little uncomfortable with our conclusions, in that there is one way in which experiments are often inferior to observational and correlational studies: They are often done in the "artificial" confines of a psychology laboratory and involve behaviors (e.g., turning wooden pegs) that seem to have little to do with the kinds of things people do in everyday life. This is one of the most common objections to social psychological experiments--they seem “artificial” and “unrealistic.” How can we generalize from such artificial situations to everyday life?

Types Of Validity

To answer this question, we need to carefully consider what it means for a study to be “real” or “valid.” Campbell and his colleagues (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979) distinguished between three different kinds of validity--internal validity, external validity, and construct validity.

Internal validity

Internal validity refers to the confidence with which researchers can draw cause-and-effect conclusions from their research results. To what extent are we certain that the independent variable, or treatment, manipulated by the experimenter is the sole source or cause of systematic variation in the dependent variable? Threats to the internal validity of research results arise when the conditions under which an experiment is conducted produce systematic sources of variance that are irrelevant to the treatment variable and not under the control of the researcher. The internal validity of a study is questioned, for instance, if groups of participants exposed to different experimental conditions are not assigned randomly and are different from each other in some important ways other than the independent variable (as in our hypothetical breakfast-eating study). It is usually easier to maintain high internal validity in a laboratory experiment because the researcher has more control over extraneous variables that might compromise the design. Even when internal validity is high, however, we need to consider how generalizable the findings are.

External validity

This term refers to the robustness of a phenomenon--the extent to which a causal relationship, once identified in a particular setting with a particular sample of research participants, can be safely assumed to generalize to other times, places, and people. When laboratory experimentation in social psychology is criticized as being “the study of the psychology of the college sophomore,” what is being called into question is the external validity of the findings. Because many laboratory experiments are conducted with college students as participants, the truth of the causal relationships we observe may be limited to that particular population (Sears, 1986). If it happens that college students--with their youth, above-average education, and nonrepresentative socioeconomic backgrounds--respond differently to our experimental treatment conditions than other types of people, then the external (but not internal) validity of our findings would be low. The same argument holds for studies conducted in one culture—how do we know that similar results would be obtained in another culture? Indeed, in an influential article, Henrich, Heine, and Norenzayan (2010) pointed out that most psychology studies have been conducted with “WEIRD” samples (people who live in Western, Educated, Industrialized, Rich, and Democratic societies), and questioned how generalizable the results are to other populations.

The issue is actually a little more subtle. No one would deny that Yale students might respond differently to a particular experimental treatment than would a sample of 50-year-old working-class immigrants or college students in another culture. External validity refers to the extent to which a particular causal relationship is robust across populations, cultures, and settings. Thus, if we were interested in inducing cognitive dissonance, we might have to use different techniques in different populations. Writing an essay in favor of the New Haven police aroused cognitive dissonance in Yale students in the 1950s, but may have been less likely to do so among conservative residents of New Haven. We would need to find another way of inducing dissonance in that second sample, perhaps by asking them to write an essay in favor of the Yale students who were protesting at the time. But even then, we would need to ask whether writing a counterattitudinal essay would have the same effects on attitude change in the two samples, or whether there are class or cultural differences in dissonance processes. For example, there is an extensive literature on whether cognitive dissonance is experienced and reduced in the same way in Western versus East Asian cultures (e.g., Heine & Lehman, 1997; Kitayama, Tompson, & Chua, 2014).

External validity concerns settings as well as samples. How do we know whether the results we find in one situation (e.g., a psychology laboratory) will generalize to another situation (e.g., everyday life)? For example, Milgram's (1974) initial studies of obedience were conducted in a research laboratory at Yale University. A legitimate question is the extent to which his findings would generalize to other settings. Because participants were drawn from outside the university and because many had no previous experience with college, the prestige and respect associated with a research laboratory at Yale may have made the participants more susceptible to the experimenter's demands for compliance than they would have been in other settings. To address this issue Milgram undertook a replication of his experiment in a different physical setting. By moving the research operation to a "seedy" office in the industrial town of Bridgeport, Connecticut, and adopting a fictitious identity as a psychological research firm, Milgram hoped to minimize the reputational factors inherent in the Yale setting. The Bridgeport replication resulted in slightly lower but still dramatic rates of compliance to the experimenter, compared to the original study. Thus, the setting could be identified as a contributing--but not crucial factor—to the basic findings of the research. Another question is whether Milgram's results were limited to the time in which the studies were conducted, when participants may have been more compliant. To find out, Burger (2009) replicated Milgram's procedures 45 years later. For ethical reasons he stopped the study after participants had delivered shocks (supposedly) of 150 volts, rather than seeing if they would go all the way to the 450-volt level at the end of the shock board. He found very similar levels of obedience up to this point and noted that in Milgram's study, most participants who delivered 150 volts continued all the way up to 450 volts.

Construct validity

To question the external validity of a particular finding is not to deny that a cause-and-effect relationship has been demonstrated in the given research study, but rather to express doubt that the same effect could be demonstrated under different circumstances or with different participants. Similarly, concerns with *construct validity* do not challenge the fact of an empirical relationship between an experimentally manipulated variable and the dependent measure, but rather question how that fact is to be interpreted in conceptual terms. Construct validity refers to the correct identification of the nature of the independent and dependent variables and the underlying relationship between them. To what extent do the operations and measures embodied in the experimental procedures of a particular study reflect the theoretical concepts that gave rise to the research in the first place? Threats to construct validity derive from errors of measurement, misspecification of research operations, and, in general, the complexity of experimental treatments and measures. One of the most difficult parts of experimental design is constructing a concrete independent variable (e.g., writing a counterattitudinal essay) that is a good instantiation of the conceptual variable (cognitive dissonance). This is essentially an issue of construct validity: How well does the independent variable capture the conceptual variable? Indeed, in the history of research on cognitive dissonance, multiple alternatives have been offered and tested (see Cooper, 2007 for a review).

The same issue holds for the dependent variable. When we devise an elaborate rationale for inducing our participants to express their attitudes toward the experiment or toward some social object in the form of ratings on a structured questionnaire, how can we be sure that these responses reflect the effect variable of conceptual interest rather than (or in addition to) the myriad of other complex decision rules our participants may bring to bear in making such ratings? And how do we know that the functional relationships observed between treatment and effect, under a particular set of operations, represent the conceptual processes of interest? This issue, we note, is relevant to the use of big data to address conceptual questions. On the one hand, such data can yield fascinating information about real-world phenomena, such as how often people tweet a particular phrase or use certain terms in web searches. The psychological constructs underlying these behaviors, however, are not always clear and require some guesswork on the part of researchers. For example, Goldy, Jones, and Piff (2022) examined Twitter data from people in the path of the 2017 solar eclipse and found that they tweeted more words synonymous with awe during the eclipse than they did before or after. This result is consistent with the idea that eclipses are awe-inspiring, though as the authors note, counting words in tweets is an imperfect measure of psychological experience.

We can now see that the experimenter is faced with a daunting task: designing a study that is well-controlled (high in internal validity), includes independent and dependent variables that are good reflections of the conceptual variables of interest (high in construct validity), and is generalizable to other settings and people (high in external validity). Internal validity may be considered a property of a single experimental study. With sufficient knowledge of the conditions under which an experiment has been conducted, of the procedures associated with the assignment of participants, and of experimenter behavior, we should be able to assess whether the results of that study are internally valid.

Issues involving construct validity and external validity, on the other hand, are more complicated. Researchers do the best they can in devising independent and dependent variables that capture the conceptual variables perfectly. But how can external validity be maximized? How can researchers increase the likelihood that the results of the study are generalizable across people and settings? One way is to make the setting as realistic as possible, which is, after all, one point of field research: to increase the extent to which the findings can be applied to everyday life, by conducting the study in real-life settings. The issue of realism, however, is not straightforward. There are several different types of realism with different implications.

Mundane Realism Versus Experimental Realism Versus Psychological Realism

As we noted, one of the most common criticisms of the laboratory experiment is that the setting is “artificial,” unlike anything people would encounter in everyday life. People know they are in an experiment and are often asked to do things (e.g., turn wooden pegs, write counterattitudinal essays) that they would rarely, if ever, do in their real lives. In short, the setting and tasks are not very realistic. But what does this mean, exactly? Aronson and Carlsmith (1968) distinguished between two kinds of realism. One, called mundane realism, refers to the extent to which events occurring in the research setting are likely to occur in the normal course of the participants' lives, that is, in the "real world." The other, called experimental realism, refers to the extent to which the situation is involving to the participants--whether they take it seriously and whether it has an impact on them.

Mundane realism and experimental realism are not polar concepts; a particular technique may be high on both mundane realism and experimental realism, low on both, or high on one and low on the other. Perhaps the difference between experimental and mundane realism can be clarified by citing an example. To study the effects of social rejection, many studies have used the Cyberball procedure, in which a participant sits at a computer and plays an online game of "catch" with two other online participants (Williams, Cheung, & Choi, 2000, Williams & Sommer, 1997). The participants throw a simulated ball back and forth until, after a few trials, the two other players exclude the participant, throwing the ball only to each other. The only real participant in the study is the one who is excluded; the other two are simulated by the computer program. For most participants, this procedure has a good deal of experimental realism. They believe they are being excluded by two other real people, and they find the experience to be intensely unpleasant and suffer marked cognitive deficits. We may assume that they are reacting to a situation that is as 'real" for them as any of their ordinary experiences. However, the experiment was hardly realistic in the mundane sense. It is uncommon in everyday life for people to play a video game with complete strangers who deliberately exclude them, especially in the context of a psychology experiment. Nonetheless, the experience is upsetting for most people.

More important than mundane realism, perhaps, is a third type of realism that Aronson, Wilson, and Akert (1994) termed *psychological realism*. This is the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life. An experiment might be nothing like what people encounter in everyday life (low in mundane realism), but still be high in psychological realism, if the psychological processes that occur are like those that occur in everyday life. The Cyberball procedure is low in mundane realism, as we saw, but the psychological processes that were triggered--feeling excluded by others for no apparent reason--were presumably like what often occurs in everyday life.

There is some overlap between experimental and psychological realism in that many of the psychological processes of interest to psychologists are ones that occur when people are reacting to impactful events in their environments. The situations in everyday life in which cognitive dissonance, prejudice, or social rejection occur are usually ones in which people are quite engaged, thus when studying these phenomena, it is imperative to devise experimental settings that are equally impactful. Such studies would be high in both experimental and psychological realism, although not necessarily high in mundane realism. Consider the counterattitudinal dissonance study: People are rarely, if ever, asked to write an essay advocating a position they disagree with; thus, these studies are low on mundane realism. But it is involving to write such an essay; indeed, when people do so for low justification, they experience the negative affective and motivational state that accompanies cognitive dissonance. Thus, the studies are high in experimental realism. And, the studies presumably capture dissonance processes that occur in everyday life, when, for example, we tell a close friend that we agree with their position on animal welfare when we really don't. If so, the studies are high on psychological realism.

In other cases, studies can be low on experimental realism but high on psychological realism. Such is the case when researchers are interested in processes that occur when people are not actively engaged or motivated to process information carefully. Examples include the study of automatic processing (e.g., Gilbert & Hixon, 1991) and investigations of peripheral or heuristic processing of persuasive messages (Chaiken, 1987, Petty & Cacioppo, 1986). As another example, psychologists are increasingly interested in the psychological causes and consequences of boredom (e.g., Westgate & Wilson, 2018). One way to study boredom is to devise tasks that are exceedingly simple to perform and that lack meaning for participants; and as such, are low in experimental realism. But, if people react to the tasks in the same way as they do to boring tasks in everyday life, the studies are high in psychological realism.

In short, experiments should always be high in psychological realism, triggering the psychological processes that the researchers are investigating. Otherwise, researchers might draw misleading conclusions. For example, some have argued for the use of scenario studies, whereby participants are asked to imagine that they were in a particular situation and to guess how they would respond (Kelman, 1966). The problem is, participants might be reluctant to report how they would respond, or genuinely not know (Nisbett & Wilson 1977). If Milgram had simply described his set up to participants and asked them how much shock they would deliver to the "learner," it is unlikely they would know (or admit) how much shock they would give. Such a study would be low in psychological realism, and provide misleading results about people's susceptibility to obedience to authority.

For a more recent example, dozens of studies have studied moral reasoning by presenting participants with the trolley scenario: A runaway trolley is barreling down the tracks, and if left unchecked, will kill five people in its path. If participants so choose, they can flick a switch that will send the trolley down another path and kill only one pedestrian. The conditions under which people will adopt a utilitarian approach (saying they would flick the switch, thereby minimizing the number of deaths), versus a deontological approach (not flicking the switch, to avoid directly harming someone) has been studied in dozens of experiments (Greene, 2013). Are these studies high in psychological realism? If the goal is to study people's moral *theories*--what they consciously believe to be morally justifiable in different situations--then the answer is yes. The studies present people with various scenarios, and measure their theories by asking them what they think is morally justifiable to do. But if the goal to assess how people *really would* respond when confronted with a dilemma like the runaway trolley, the studies are decidedly low in psychological realism. The psychological processes triggered when suddenly noticing a real runaway trolley and having only a moment to decide what to do are likely different from those triggered when reading about a hypothetical trolley and imagining how one should respond. Indeed, one study placed participants in a real trolley-like situation, telling them that they had 20 seconds to decide whether to allow five mice to receive a painful electric shock or push a button that would redirect the shock to a single mouse (Bostyn et al., 2018; note that although participants believed the shocks would be delivered, no mice were actually shocked). The percentage of people who pressed the button in this real situation was higher than the percentage of participants who, after reading a description of the study, *said* they would press the button (84% vs. 66%). These results suggest that studies presenting participants with the trolley scenario may be low in psychological realism, if the goal is to assess how they respond to actual moral dilemmas in their everyday lives.

Should studies always be high in experimental realism? It is important that participants be engaged enough that they take the study seriously and pay attention to the instructions, but *how* engaged they should be depends on the nature of the psychological phenomenon under study. To study some phenomena, such as cognitive dissonance or conformity pressures, participants must be put in situations that are involving enough to trigger these processes. But as we noted, sometimes researchers are interested in phenomena that involve little engagement, such as boredom, and here experimental realism is deliberately lower. Note that of the different kinds of realism we have discussed, mundane realism is the least important. Indeed, sometimes it is necessary to create an artificial setting to gain enough control over the situation to maintain high internal validity. In fact, as we will see next, creating a setting that is like real life might not be important at all.

External Validity: Is It Always A Goal?

It is often assumed that all studies should be as high as possible in external validity, in the sense that we should be able to generalize the results as much as possible across populations and settings and time. Sometimes, however, the goal of the researcher is different. Mook (1983) published a provocative article entitled, "In defense of external invalidity," in which he argued that the goal of many experiments is to test a theory, not to establish external validity. Theory testing can take a variety of forms, some of which have little to do with how much the results can be generalized. For example, a researcher might construct a situation in which a specific set of results should occur if one theory is correct, but not if another is correct. This situation may be completely unlike any that people encounter in everyday life, and yet, the study can provide an interesting test of the two theories.

Mook (1983) gives the example of Harlow's classic study of emotional development in rhesus monkeys (Harlow & Zimmerman, 1958). Infant monkeys were separated from their mothers and placed in cages with wire-covered contraptions that resembled adult monkeys. Some of the wire monkeys were covered with terry cloth and were warmed with a light bulb, whereas others were bare and uninviting. Nourishment (in the form of a baby bottle) was sometimes available from one type of monkey and sometimes from the other. Harlow found that the monkeys clung to the terry cloth "mother," regardless of whether it contained the bottle of milk. These results were damaging to drive-reduction theories that argued that the monkeys should prefer nourishment over emotional comfort. Was this study high in external validity? Clearly not. There was no attempt to randomly select the monkeys from those reared in the wild, or to simulate conditions that monkeys encounter in real-life settings. Nonetheless, if theories of drive reduction that were prevalent at the time were correct, the monkeys should have preferred the nourishment, regardless of which "monkey" it came from. The researchers succeeded in devising a situation in which a specific set of actions should have occurred if a particular theory was right--even though the situation was not one that would be found in everyday life.

Mook also pointed out that some experiments are valuable because they answer questions about "what can happen," even if they say little about "what does happen" in everyday life. Consider Milgram's experiments on obedience to authority. There was little attempt to simulate any kind of real-life setting in these studies; people are never asked to deliver electric shocks to a stranger who is performing poorly on a memory test. The results were informative, however, because it was so surprising that people would act the way they did under *any* circumstances. This is sometimes referred to as "proof of principle." The fact that people *can* be made to harm a complete stranger, because an authority figure tells them to, is fascinating (and frightening) despite the artificiality of the setting.

Mook's (1983) position is persuasive, and we heartily agree that the goal of many experiments is to test a theory, rather than to establish external validity. Nonetheless, we believe that even if external validity is not the main goal of study, it should never be completely forgotten. The importance of a theory, after all, depends on its applicability to everyday life. The reason Harlow's study is so important is because the theories it addresses--drive-reduction and emotional attachment--apply to humans as well as monkeys and to many situations beyond cages and wire mothers. It is precisely because the *theories* are generalizable (i.e., applicable to many populations and settings) that a test of those theories is important. Thus, a specific study might test a theory in an artificial setting that is low in external validity, but why would we conduct such a study if we didn't believe that the theory was generalizable? Similarly, Milgram's results are so compelling because we can generate important, real-life examples of times when similar processes occurred. Indeed, the inspiration for Milgram's study was the Holocaust, in which seemingly normal individuals (e.g., guards at prison camps) followed the orders of authority figures to the point of committing horrific acts. Thus, if we were to conclude that the psychological processes Milgram uncovered never occur in everyday life, we could justifiably dismiss his findings. The fact that these processes appear to be like those that occurred at some of humankind's darkest moments--such as the Holocaust--is what makes his results so compelling.

We are essentially reiterating the importance of psychological realism in experimentation. To test a theory, it may be necessary to construct a situation that is extremely artificial and low in mundane realism. As long as it triggers the same psychological processes as occur outside of the laboratory, however, it can be generalized to those real-life situations in which the same psychological processes occur. Of course, as discussed earlier, claims about psychological realism cannot be taken completely on faith; only by replicating a study in a variety of settings can external validity be firmly established.

The Basic Dilemma Of The Social Psychologist

It should be clear by now that the perfect social psychology study would be experimental instead of correlational, extremely high in psychological realism, and study the psychological processes underlying an important phenomenon. Ideally, the study would be conducted in a naturalistic setting in which participants were randomly assigned to experimental conditions and all extraneous variables were controlled. Unfortunately, it is next to impossible to design an experiment that meets all these demands. Indeed, almost no study ever has. One of the few exceptions, perhaps, is the Lepper, Greene, and Nisbett (1973) classic study of the overjustification effect, which was conducted in a naturalistic setting (a preschool) in which participants (3-and 4-year old children) were randomly assigned to various conditions of rewards or no rewards for drawing with felt-tip pens, and the dependent variable was how much the kids played with the pens 2 weeks later during a normal classroom activity. (An interesting social psychological party game is to see if you can come up with any other studies that meet all the conditions we have laid out for the Platonic Social Psychological Experiment--there are not many.) Aronson and Carlsmith (1968) called this the basic dilemma of the experimental social psychologist. On the one hand, we want maximal control over the independent variable, to maintain internal validity. But, by maximizing internal validity, we often reduce external validity (e.g., by conducting our study in the lab instead of the field).

A solution to the basic dilemma of the social psychologist is to not try to "do it all" in one experiment. Instead, a programmatic series of studies can be conducted in which different experimental procedures are used, in different settings, to explore the same conceptual relationship. It is in this realm of conceptual replication with different scenarios that the interplay between laboratory and field experimentation is most clear. However, in considering these interrelationships, the tradeoff mentioned earlier between control and impact in different experimental settings becomes especially salient. To be defensible, weaknesses in one aspect of experimental design must be offset by strengths or advantages in other features, or the whole research effort is called into question. This dictum is particularly applicable to field experiments in which inevitable increases in cost and effort are frequently accompanied by decreases in precision and control that can be justified only if there are corresponding gains in construct validity, impact, or the generalizability of findings.

Multiple Instantiations Of The Independent Variable

Essentially, there are two properties that we demand of a series of experiments before we are convinced that we understand what the conceptual interpretation should be. First, we ask for a number of empirical techniques that differ in as many ways as possible, having in common only our basic conceptual variable. If all these techniques yield the same result, then we become more and more convinced that the underlying variable that all techniques have in common is, in fact, the variable that is producing the results. For example, the contact hypothesis--the idea that intergroup contact can, under the right circumstances, reduce prejudice--has been demonstrated with several different ways of operationalizing contact, including actual and imagined interactions between group members (Miles & Crisp, 2014; Paluck, Green, & Green, 2019; Zhou et al., 2019).

Multiple Instantiations Of The Dependent Variable

Second, we must show that a particular empirical realization of our independent variable produces many different outcomes, all theoretically tied to the independent variable. For example, researchers testing the contact hypothesis have included several outcome measures, including self-reported feelings about the outgroup, overt behavior toward the outgroup, and measures of implicit attitudes (Paluck et al., 2019).

Replications

Programmatic research inherently involves replications, whereby researchers duplicate and extend their findings. For instance, in a replication of Cohen's counterattitudinal essay study, Linder, Cooper, and Jones (1967) found the same results when participants were made to feel they had a choice about whether to write the essay (as they were in Cohen's study), but not when participants were simply instructed to write the essay. Performing a counterattitudinal behavior causes dissonance, they concluded, only when participants feel they freely chose to perform the behavior. In other words, they not only successfully replicated the counterattitudinal dissonance finding obtained by Cohen, but also made a significant conceptual advance, by showing that the standard dissonance effect occurs only when participants are made to feel that they freely chose to write the essay. This is the way experimental science progresses--by demonstrating the conditions under which a phenomenon does and does not occur.

In recent years, the issue of replicability has moved to the forefront of the field, with considerable debate over the extent to which studies in social psychology, and the field of psychology more generally, can be replicated (e.g., Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011). We do not have the space to present a comprehensive review of the "replication movement," but we do offer some observations, both here and later in the chapter, as they relate to the methodological issues surrounding experimentation (see Giner-Sorolla, 2025, for a more general discussion of meta-science).

We note that there has been considerable debate over how to interpret the results of some high-profile replication projects, such as the Reproducibility Project: Psychology in which 36 out of 100 replication attempts yielded significant results (Open Science Collaboration, 2015). As with all replications, it can be difficult to determine the correct metric for assessing replicability, whether error in measurement is adequately accounted for, and whether replications faithfully duplicate the methods of the original studies (e.g., Gilbert et al., 2016). In short, the same criteria we use in assessing original experimental designs apply to replication studies--namely, the extent to which they are high in internal, psychological, and external validity. The latter is particularly important if the goal is estimation of the replicability of a given field as a whole. Doing so would require a random sampling of all studies in the field, which poses a considerable logistical challenge. Indeed, most large-scale replication projects have been explicit about the kinds of studies they excluded (e.g., ones that could not be conducted online, field studies, studies with nonstudent participants, etc.). It would be highly impractical--perhaps impossible--to randomly select studies from the field at large, which makes sweeping statements about the replicability of entire fields difficult.

Finally, we note that it is unclear what that “replicability success rate” should be for a healthy science. Many people assume that it should be 100%, but as pointed out by B. Wilson and Wixted (2018), that could be achieved only by adopting a very conservative approach to science that minimized Type 1 errors (false positives) at the expense of an increase in Type II errors (false negatives, or failing to discover true effects). That is, it would encourage researchers to conduct “safe” studies that made only incremental advances, and discourage them from taking risks by exploring important new phenomena that might lead to blind (nonreplicable) alleys. Indeed, just as simulations suggest that scientific discovery is maximized by running a larger number of moderately powered studies, rather than concentrating resources on a single large well-powered one (Finkel, Eastwick, & Reis, 2017), we suggest that seeking a balance between false positives and false negatives is likely to be the most fruitful path forward (Finkel, Eastwick, & Reis, 2015).

Now that we have discussed the value of experiments, we turn to a discussion of how to conduct one. In discussing the nuts and bolts of experimentation we will not lose sight of the important questions about the advantages and disadvantages of experiments and will return to these issues frequently.

PLANNING AND CONDUCTING AN EXPERIMENT

The best way to describe how to conduct an experiment is to take a real study and dissect it carefully, examining how it was done and why it was done that way. We have chosen for illustrative purposes a classic study on bystander intervention conducted by Darley and Latané (1968). The study was inspired by a highly publicized, real-life tragedy, namely the murder of Kitty Genovese in the Queens section of New York City in 1964. When she returned home to her apartment late one night, Ms. Genovese was brutally murdered in an assault that took a full 45 minutes. The New York Times later estimated that 38 of Ms. Genovese’s neighbors saw or heard the attack, but that not one of them tried to help, not even by calling the police. We now know that this account was probably wrong, and that some neighbors did try to help (Roberts, 2020). Nonetheless, the story of the “apathetic neighbors” received widespread attention, prompting editorials about the heartlessness of New Yorkers and the dehumanizing consequences of urban life. The neighbors were widely castigated for their seeming callousness and lack of empathy.

Two social psychologists living in New York at the time, John Darley and Bibb Latané, wondered whether there was a different explanation--one that was in many ways more frightening. Maybe the neighbors' lack of response had more to do with the conditions under which the murder occurred, which would have made anyone unlikely to help. The very fact that so many people witnessed the murder--and knew that they were one of many witnesses--may have made it less likely that any one of them would feel that it was "on their shoulders" to intervene. Latané and Darley (1968) referred to this as the diffusion of responsibility--the more people who are present during an emergency, the less any one of them feels that it is their responsibility to help.

That's an intriguing hypothesis, but how might one go about testing it empirically? Note that it is a causal hypothesis: The number of bystanders determines how likely any one of them is to help. Thus, the best way to test it is to perform an experiment. To do so, a researcher would need to stage an emergency of some sort, vary (randomly, of course) how many people witnessed that emergency, and then observe how many people tried to help. But what kind of emergency would be feasible to stage? (Surely not a murder.) And how could we make sure that everything in the study was held constant, except the number of bystanders (that is, that internal validity would be high)?

Let's see how Darley and Latané went about it. They recruited 72 New York University students, each of whom took part individually. When the participant arrived, they were ushered to a long corridor with several rooms on either side and seated at a table in one of the rooms. After filling out a background questionnaire, they were asked to put on a pair of headphones with a microphone, through which all subsequent instructions were given. The experimenter left the room and explained, over the intercom, that the purpose of the study was to learn about the kinds of personal problems college students face by having them engage in a frank discussion with other students. Because of the sensitive nature of the topic, he continued, several precautions would be taken. First, the participants would be anonymous, and in fact would not meet face-to-face, but communicate solely over the intercom system. Second, the experimenter would not listen in on the conversation, so that the students would not feel inhibited by his presence. Third, the discussion would follow an organized pattern, whereby each person would discuss any personal problems they were having for 2 minutes, after which everyone would have 2 minutes to react to what the others had said. Only one person's microphone would be on a time, to maintain order in the conversation.

Already we can see that Darley and Latané created a quite artificial situation, that is, one that was very low on mundane realism. In everyday life, people rarely, if ever, have group discussions over an intercom system. Why not have people talk face-to-face, to be more like everyday life? One reason was to be able to experimentally manipulate how many people participants thought were present in the group. Participants were randomly assigned to conditions in which they heard one, two, or five other participants take part in the discussion. To make sure that everything else about the experience was held constant, the researchers pre-recorded the comments made by the "other participants." That is, there was only one real participant in the study, though they believed that they were listening to a live discussion.

Further, pre-recording the voices allowed the researchers to stage an identical emergency for all participants. In every condition, the first person to speak mentioned that he sometimes had seizures, especially when under stress. Each of the other group members, including the real participant, then talked about any problems they were having adjusting to college. When it was the first participant's turn to speak again and react to what the others had said, he started out calmly, but then began to stutter and talk incoherently. It soon became clear that he was having one of the seizures he had previously mentioned. He gasped and repeatedly asked for help from the others, finally yelling, “I'm gonna die-er-er-I'm . . . gonna die-er-help-er-er-seizure-er”--after which his 2 minutes were up and his microphone automatically shut off (Darley & Latané, 1968, p. 379). Again, crucially, all participants witnessed the exact same emergency--the only thing that varied was how many others they thought were witnessing it as well.

Did the (perceived) size of the group influence the likelihood that participants would help the poor fellow having the seizure? It did: by the time the victim’s microphone was shut off, 85% of the people who thought they were the only person to hear the seizure had come out of their cubicle to try to help, by finding the victim or the experimenter. This percentage dropped to 62% in the condition in which participants thought there was one other bystander, and to 31% in the condition in which they thought there were four other bystanders. The researchers continued to monitor helping for up to 6 minutes after the seizure began, at which point they terminated the study and fully debriefed the participants. By that time, 100% of those in the two-person group condition had helped, whereas only about 80% and 60% had helped in the three-person and six-person groups, respectively.

On the one hand, this was a very “clean” test of the bystander hypothesis, one that convincingly showed that, at least in the situation the researchers constructed, that the number of bystanders had a large effect on the likelihood that individual participants would try to help. Indeed, this study has become one of the most famous in all social psychology. But we hope you can see the limitations of the study and the trade-offs it involved. The researchers went to some lengths to maintain high internal validity, by conducting the study in the laboratory and placing participants in an artificial situation. This allowed them to be confident that it was the independent variable (the number of bystanders) that exerted a causal effect on helping variable, and not something else. But this came at a cost, namely observing helping in a situation that was far removed from anything people experience in everyday life. This is the basic dilemma of the social psychologist we discussed earlier: By maintaining high internal validity, the researchers sacrificed external validity. Think, for example, how different the Darley and Latané experiment was from the real-life incident that inspired it--the murder of Kitty Genovese. The two situations could hardly be more different; in one, college students participating in a psychology study heard another student have a seizure over an intercom; in the other, neighbors were awakened late at night by desperate cries for help from a woman being brutally attacked. Can we really learn anything about why Kitty Genovese’s neighbors (supposedly) failed to help?

The short answer is no, given how different the situations were. But this wasn’t really the point of Darley and Latané’s study. Rather, they were testing a hypothesis about a psychological process thought to occur in real emergencies of many types, namely the diffusion of responsibility that occurs when there are multiple witnesses. Psychological realism was, arguably, quite high in their study, in that participants believed that a real emergency had occurred and were visibly shaken by it. That is, even though the setting in which the emergency occurred was unlikely anything participants would experience in everyday life, the psychological processes that were triggered were probably the same. Here’s how Darley and Latané put it:

> On the one hand, subjects worried about the guilt and shame they would feel if they did not help the person in distress. On the other hand, they were concerned not to make fools of themselves by overreacting, not to ruin the ongoing experiment by leaving their intercom, and not to destroy the anonymous nature of the situation which the experimenter had earlier stressed as important. For subjects in the two-person condition, the obvious distress

> of the victim and his need for help were so important that their conflict was easily resolved. For the subjects who knew there were other bystanders present, the cost of not helping was reduced and the conflict they were in more acute. (p. 382)

These very same psychological processes, the authors suggest, occur in those who witness real-life emergencies, maybe even Kitty Genovese's neighbors. But should we take their word for it? Of course not; a single study cannot settle such questions about external validity. Instead, as we mentioned, it is important to pursue such questions programmatically, by replicating and extending the findings in a variety of settings with diverse populations. One reason Latané and Darley's research has become so famous is that they and others conducted a programmatic series of studies that tested the effects of the number of bystanders on helping in different kinds of emergencies (e.g., potential fires, fights, and accidents; Latané & Darley, 1968; Shotland & Straw, 1976; Staub, 1974); in the field as well as the laboratory (e.g., Latané & Dabbs, 1975); with a variety of populations (e.g., children, college students, subway riders, future ministers; e.g., Darley & Batson, 1973; Latané & Nida, 1981; Plötner et al., 2015). A meta-analysis examined 105 tests of the bystander effect and found an overall effect size of g = -.33 (such that the more bystanders present, the less likely victims were to be helped)--a small to moderate effect (Fischer et al., 2011).

It is worth noting that a recent study took a different approach to studying helping in emergencies, by analyzing videos of actual conflicts recorded by surveillance cameras in three different countries (Philpot et al., 2020). The conflicts ranged from "the mildest animated disagreements to grave physical violence" (p. 69) between two or more individuals. The researchers coded whether any of the bystanders "attempted to placate the conflict" in any manner (p. 70), and found that the greater the number of bystanders, the *more* likely at least one person was to try to help. In one respect, this study is far superior to Latané and Darley's, in that it examined helping in real-life emergencies. But because of the lack of random assignment to group size, the meaning of this correlation is unclear (the number of bystanders was likely confounded with many other variables). Nor, as the authors point out, does it speak to the question of whether bystanders are more likely to help when they are the sole witness to an event compared to when there are others present.

Choosing The Type Of Experiment To Perform: Conducting Studies In The Laboratory Versus Online

Let us assume that you have a terrific hypothesis about the causes of human behavior. Following our logic up to this point, you recognize you need to do an experiment to test your hypothesis (as opposed to a correlational study). But now you need to decide the setting in which you will conduct your experiment--should it be in the laboratory, online, or in the field? We've already discussed the pros and cons of doing studies in the field, such as the gain in external validity but (often) the loss of internal validity. We refer the reader to Paluck and Shah (2025) for an excellent discussion of why field studies are important and how to do them.

Say you've decided that it's not feasible to conduct your initial study in the field. Should you begin by doing your study in the laboratory or online? The internet has made it possible to order groceries, adjust a thermostat, and even buy a house or get married from one's home, office, or wilderness retreat. It is thus not surprising that researchers have also found innovative ways to leverage this medium for psychological research. There are clear advantages to this approach. If you can devise a way to test your hypothesis on a survey platform (e.g., Qualtrics), using participants drawn from a service such as mTurk or Prolific, you can get large amounts of data in a short period of time. It is not unusual to post a survey on one of these services and have dozens of people complete the study before the day is out. You will know very soon whether your hypothesis is confirmed, as opposed to collecting your data in a laboratory setting, which can take several months.

In addition to getting your data quickly, you will likely be able to collect much larger samples than you could in a laboratory study. This is helpful, especially in detecting experimental effects, because underpowered studies increase the rate of both false negatives and false positives. When sample sizes are small, researchers are more likely both to miss effects that exist, as well as to falsely "detect" effects which do not, simply because the data are noisier (Button et al., 2013). Longitudinal studies are often easier to conduct wholly or partially online; indeed, a common hybrid paradigm combines an initial in-person laboratory experiment with an online follow-up days, weeks, or even months later. Such designs can not only increase statistical power (via repeated measures), but also allow for studying processes naturalistically as they unfold over time. And there is no doubt that online data collection has greatly expanded researchers' access to more diverse and representative populations. Many data collection services now offer researchers the ability to target even very niche sub-samples (e.g., gun owners), who may have been previously difficult to recruit.

However, we note that larger sample sizes can be offset by other challenges typically not present in laboratory settings. First, because online experiments are often--by necessity--lower in experimental realism, manipulations designed for use in online experiments may yield smaller effects than those designed for the lab. They may also be noisier: online participants may be distracted by any number of environmental factors (e.g., location, noise, the presence of others) that could be eliminated or held constant in the laboratory Such distractions can not only divert attention away from experimental procedures, but complicate measurement of context-sensitive variables, such as mood. Thus, larger online samples may at times be (partially) offset by smaller experimental effects.

Most importantly, is it feasible to test your hypothesis in an online study? By their very nature, some hypotheses are difficult or impossible to test online. These include hypotheses about social behavior that require participants to have face-to-face interactions with other people and studies in which it is necessary to observe people's behavior or measure their physiological responses. In other studies, it is necessary to put people in impactful situations in which they are active participants; they experience an unfolding series of events and must react to those events as they occur. The Darley and Latané (1968) epilepsy study was clearly an impact experiment; hearing a fellow participant have a seizure, and having to decide if and how to intervene, was highly involving, even upsetting. Another example is studies that employ the Cyberball game. Participants typically start out thinking the game will be a simple, maybe even pointless exercise; but as the other participants repeatedly fail to throw them they become much more involved and experience the sting of social rejection.

Contrast this to a typical judgment study in which participants are more like passive observers; they are asked to recognize, recall, classify, or evaluate stimulus materials presented by the experimenter. Little direct impact on participants is intended, except insofar as the stimulus materials capture people's attention and elicit meaningful judgmental responses. Many social cognition studies are of this type, in that they are designed to study how people process information about other people, by presenting them with scenarios and measuring self-reported judgments, such as hiring preferences. Which type of study should you do?

As mentioned, it depends on the psychological process you are studying. Richeson and Trawalter (2005), for example, hypothesized that interacting with someone of a different race can deplete people's cognitive resources. To test this, they randomly assigned White participants to meet in person with either a White or Black accomplice. As predicted, they found that White participants who interacted with a Black accomplice were more cognitively depleted than White participants who interacted with a White accomplice. Given that this study required participants to have an actual interaction with someone of the same or different race, it would have been impossible to do online.

Other hypotheses, however, are judgmental in nature. Ma, Axt, and Kay (2019), for example, hypothesized that feeling a lack of control over the world would increase stereotyping. To test this, they conducted an online study in which mTurk participants were randomly assigned to write about something that had happened to them recently over which they did or did not have control, then complete a questionnaire that measured gender stereotyping. As predicted, they found that those who wrote about not having control exhibited more stereotyping.

Typically, impact experiments are easier to conduct in the laboratory. But this is not a hard and fast rule. Researchers have tested the bystander effect online, for example, by having participants take part in internet discussion forums and seeing how likely they are to respond to a request from another online participant for help. As predicted, people are less likely to offer help when they think there are many other users online than when they think they are the only other person in the forum (Markey, 2000; van Bommel et al., 2016). Clearly this is a different (and less impactful) psychological situation than hearing another person have a seizure, but the findings add to our knowledge about the generalizability of the bystander effect.

Issues With Online Experimentation

Let's say you have a hypothesis that you think can be tested online. There are several issues to consider when planning such a study. Although we always advise care be paid to ensuring participants understand experimental procedures, doing so is particularly important in online studies, where the experimenter might not be present to correct misperceptions (Crump et al., 2013). Comprehension checks can be added prior to the critical manipulation, with progress in the study contingent on successful answers. Likewise, researchers can specify prerequisites for participation to ensure greater uniformity and ask participants to verify that they are participating at a time when they are alone, free of distractions (e.g., no TV, music), and able to complete the study in a single sitting. In a study featuring audio or video content, participants may first be asked to correctly report the content of sample stimuli (e.g., correctly identifying, for instance, that a sample audio track contains bird and not train noises).

Such steps not only ensure that the manipulation is clear to participants, but also help reduce participant dropout. Although dropout, or *attrition,* may appear a minor inconvenience, it can pose a serious threat to internal validity (Zhou & Fishbach, 2016). If participants are equally likely to drop out in all conditions (i.e., missing completely at random; MCAR), a researcher may end up with a smaller sample but can still be relatively confident that any changes in the dependent variable are due to the experimental manipulation. A bigger problem exists when participants are more likely to drop out of one condition than the other, or when different types of people are more likely to drop out of different conditions, often labeled *missing at random (MAR)*. In such cases, participants are no longer randomly assigned to condition, and we can no longer be confident that it is our experimental manipulations--rather than differences in the types of people in the different conditions--that are causing our effects.

Such attrition occurs often in online experiments but is sometimes overlooked. Common manipulations in social psychology such as construal level (i.e., describing an event in abstract vs concrete terms) and ego depletion have attrition rates ranging from 30%-70% (Zhou & Fishbach, 2016). This is particularly problematic when paired with control conditions in which attrition is less likely. For instance, in one online study, over 75% of participants dropped out when asked to write a 100-letter paragraph without using the letters A or N; whereas attrition was much lower (though still high at 23%) when participants were asked to write a 100-letter paragraph without the letters X or Y. Even if equal numbers dropped out in both conditions, it is no guarantee that they did so for the same reasons. Distinguishing between data that are missing completely at random (MCAR), missing at random (MAR) or missing *not* at random (MNAR) has long been a challenge in correlational and longitudinal designs; when run online where attrition is often high, it poses just as formidable a challenge for experiments.

When attrition occurs online, it may not only weaken or eliminate effects, it can even reverse them. For instance, Zhou and Fishbach (2016) asked participants to recall 12 or 4 happy events. Many more people dropped out in the 12-event than the 4-event condition, presumably because they didn’t want to exert that much effort. As a result, the remaining participants in the 12 happy events condition reported that the task was easier and less effortful than did participants in the 4 happy events condition. In another example, online participants were randomly assigned to imagine applying either eyeliner or aftershave at home, and then report their weight. Men were more likely to drop out of the eyeliner condition, whereas women were more likely to drop out of the aftershave condition. As a result, because, on average, women weigh less than men, the results suggest that imagining applying eyeliner *caused* people to weigh less. Although such results are clearly absurd, selective attrition can result in other more plausible (yet equally false) findings, which are harder to detect.

How can researchers reduce differential attrition in online experiments? One approach is to follow best practices: compensate participants fairly and keep the study short, easy, and engaging. Participants are less likely to drop out of an engaging 3-minute study than a tedious 30-minute one. Not all research questions, however, are conducive to this approach, and it can be difficult--or even inappropriate--to make all experimental conditions equally challenging or interesting. For instance, it would be impossible to study the effects of boredom experimentally if all conditions were equally boring! And pay is not always a straightforward solution; in one study, higher compensation increased how many people were willing to participate, but did not affect whether they completed the study (Göritz, 2014). Furthermore, as we know from research on cognitive dissonance, pay itself can psychologically influence participants.

In such cases, Zhou and Fishbach (2016) suggest an alternate approach: reducing differential attrition by warning participants in advance of any aversive elements the study will entail that might entice participants to drop out, appealing to their conscience by explaining why that would be a problem, and securing their agreement to complete the entire task, “signed” by an identifier, such as their Worker ID. Such pre-commitments were successful in cutting attrition almost in half on a standard construal level manipulation. Although this solution reduces generalizability (due to self-selection in who chooses to participate), it preserves random assignment by ensuring that all participants who enter the study are equally likely to be randomly assigned to—and stay in their—assigned conditions. Another solution, employed by Pfattheicher et al. (2021), sidesteps the issue by using a within-person design: all participants participate in all conditions, with order of condition randomly assigned. Only the first (between-participants) comparison is analyzed, and only for those participants who completed the entire study.

With this procedure, researchers ensure that participants predisposed to drop out of one condition are dropped equally from both conditions. For example, participants are more likely to stop watching a boring than an interesting video, and this effect is stronger for some people than for others. In a traditional between-subjects design, more participants would drop out of the boring video condition than the interesting video condition, resulting in differential attrition and the problems that entails. By ensuring that participants in the interesting video condition must also subsequently watch the boring video to be retained in analyses, participants disinclined to watch boring videos will be dropped from both the interesting and boring video conditions (rather than exclusively from the boring video condition). Such solutions minimize the possibility that attrition due to third variables (e.g., boredom proneness, gender) is driving observed differences between conditions, rather than the experimental conditions themselves. Although such precautions may still result in high attrition and resulting problems of generalizability, they sidestep the problems with causal inference that differential attrition can cause.

Participants in online studies are often self-selected, which can limit generalizability in ways similar to those seen in laboratory studies (that recruit, for instance, non-representative undergraduate samples). Project Implicit (https://implicit.harvard.edu/implicit/), for instance, offers visitors the opportunity to test themselves for implicit bias and thus attracts samples interested in doing so, which skew younger, more educated, more politically liberal, and, often, less biased than the national average (as compared, for instance, to Pew or Gallup samples). Media reports, which rarely affect data collection in laboratories, can also skew online samples in surprising ways. For instance, in the summer of 2021 researchers began observing a massive gender imbalance on Prolific, with samples often skewing as high as 80% women. It turned out that a popular TikTok video promoting Prolific as a “side hustle” had gone viral among young women; the video was viewed over four million times and led to a massive increase in the number of young women in their 20s participating in online studies (Letzter, 2021). Similar media effects have also been observed on Project Implicit and other public research sites following press coverage. As in any study, collecting and checking participant demographics prior to data analysis can help detect such problems, and may be particularly important in online data collection, where less is known about one’s participants.

Demographic and comprehension checks not only ensure that participants understand the experimental procedures, in online studies they have a secondary purpose: the detection and elimination of “bots,” or non-human participants. Concerns about data quality in online studies often center on whether participants are real people participating in good faith. Automated scripts, colloquially called “bots,” can be quite sophisticated; by quickly filling out large numbers of responses, they generate revenue for their creators and headaches for researchers. Such scripts can be run fully on their own, or be assisted by humans who fill in checks meant to detect them. Data “farms” may likewise outsource fake accounts to real people, often underpaid, who are tasked with quickly completing as many studies as possible. Finally, even “real” participants may participate haphazardly, or even knowingly respond dishonestly (e.g., to qualify for study inclusion).

Researchers can take steps to help ensure data quality in their online experiments. In addition to the practices outlined above, many online participant platforms periodically prepare reports on data quality in their own participant pools and seek to deter instances of bots or farming. For instance, participants may be required to provide identifiers to prevent false accounts and providers may limit new accounts from previously problematic addresses. Furthermore, many platforms track metrics that allow researchers to screen participants based on prior data quality and participation history (although not all IRBs permit researchers to make full use of these features). Such options may be unavailable outside of commercial platforms, such as university participant pools or community samples. In such cases, researchers may need to adopt their own data quality measures.

For instance, Qualtrics and other online survey platforms offer researchers the ability to include Captchas (to detect automated responses) and to screen IP-addresses (and commonly used VPNs) to prevent multiple submissions. Other precautions include comprehension checks (as described above), which participants must successfully pass to participate in the study, or disguising key selection criteria in the prescreen by embedding them with other questions, with the goal of ensuring genuine responses. Steps can also be taken to prevent low-quality responding and/or cheating during the study itself. Timing, which allows researchers to track how long participants view a page and/or require that they spend a minimum time doing so, can reduce speeding or limit the total time available to complete the study—and can be used to detect instances where participants may be away from the keyboard. More complex studies may benefit from teleconferencing, in which an experimenter interacts directly with the participant, over online video platforms such as Zoom. And, increasingly, commercial platforms such as Facebook are engaging in online experiments in which their users may not even be aware they are participants (which can raise its own ethical issues).

Timing and similar precautions provide data that researchers can use to report and screen low-quality responses as part of data cleaning and analysis. Computer-generated responses can be distinguished from human-generated ones with a wide variety of statistical techniques (see Dupois et al., 2019; Meade & Craig, 2012 for a comparison of such indices); open-ended questions can be particularly helpful for detecting non-human or low-effort responses. Perhaps the most common (and controversial) such measure are attention checks. Attention checks can range from simple face-valid items (e.g., asking if participants were paying attention or following instructions), to memory checks (e.g., asking participants to recall instructions from earlier in the study) and hidden questions (e.g., instructions to ignore the actual item text and select a specific number in response to a question).

While the appeal of attention checks may seem obvious, and their use straightforward, there are several issues to consider before using them. First, attention checks may not always measure what researchers think they do. For instance, participants may miss trap questions not because they weren't paying attention, but because of poor reading comprehension, which may be confounded with variables such as SES, education, and age. Second, the inclusion of attention checks may itself alter the behavior of participants by making them more cautious or alert to "sneaky" behavior on the part of the researchers. For instance, in one study, trap questions (e.g., "Ignore the instructions above and select the option for 'other' to show you are paying attention") made participants more cautious and deliberative on subsequent tasks (Hauser & Schwartz, 2015). Thus, while placing attention checks early in a study can help screen out inattentive responders prior to random assignment, doing so can also inadvertently affect participants' later responses. On the other hand, conditioning inclusion on such checks after the manipulation runs the same risks to internal validity as differential attrition and internal analyses, discussed earlier. Factual manipulation checks (in which participants are asked to recall instructions from earlier in the study) run a lower risk of altering participants' behavior, especially if they are placed at the end of the study. But, they may not accurately reflect participants' understanding of the critical manipulations that came earlier in the study.

In general, when considering how to handle low-quality responses, it is preferable that analytic decisions focus on exclusion criteria administered prior to random assignment. Excluding participants based on post-manipulation measures (including attention checks) reduces the ability to conclude, definitively, that it is the manipulation and *only* the manipulation driving any observed differences in results. One solution is to deploy attention and other data quality checks as robustness analyses--to ensure that results remain consistent, even when suspicious or low-quality responses are eliminated.

Given these issues with online experimentation, how should you test your hypothesis? We note that when proper procedures are followed, data quality in online studies can be very good. Indeed, several studies have compared the same design conducted in the laboratory (using traditional samples) versus online in just this way, and found results to be virtually identical (e.g., Germine et al., 2012; Hilbig, 2016), even with young children (Nelson et al., 2021; Nussenbaum et al., 2020). As we have noted, however, this will surely depend on the kind of study you want to conduct; many high-impact studies involving social interaction and behavioral measures simply can't be conducted online.

Perhaps the cleanest way to handle issues with selective attrition and data quality in online studies is to conduct online experiments in the context of programmatic research, which includes in-lab replications of key effects. If effects are consistent both online and in the lab, researchers can be confident that results are not due to quirks of the research setting. For example, Furrer, Wilson, and Gilbert (in press) hypothesized that there is an *illusion of unfairness*, whereby people find random events to be less fair if other people appear to have controlled them. Suppose, for example, that an election has resulted in a tie and that a coin flip is used to determine the winner. It shouldn't matter who gets to call heads or tails and flip the coin, because coin flips are random. Furrer et al. (in press), however, thought it would matter, namely that, in this example, the candidate who loses would think the outcome is less fair if the other candidate got to call heads/tails and flip the coin. To test this hypothesis, Furrer et al. conducted three types of studies: (a) Highly controlled laboratory studies, in which a coin flip determined whether participants received a negative or positive reward; (b) online studies using mTurk participants, who thought they were competing with another mTurk participant for positive or negative outcomes, determined by a virtual coin flip; and (c) field studies, in which two passersby competed for a snack, determined by a coin flip. In all three settings, the illusion of unfairness hypothesis was supported: participants who lost thought the outcome was less fair if the other person had been the one to call heads/tails and flip the coin. Consistent with our earlier point that online studies can be "nosier," however, the effect size of the basic illusion of unfairness result was larger in the laboratory studies (mean d = .52) than in the online studies (mean d = .32).

Each of these studies had advantages and disadvantages. The laboratory studies were quite time consuming and had smaller samples, and participants were limited to college students. But they also had the most experimental control. The mTurk studies had samples that were 10 times as large but were undoubtedly "noisier." Participants in the field study thought they were simply doing a consumer survey, and thus may have been less wary of being in an experiment. But they were sometimes distracted by what was going on around them. The fact that the illusion of control was found in all three types of studies "cancels out" these disadvantages and makes us more confident in the results.

As this example illustrates, it is possible to do some kinds of high-impact studies online. But some probably shouldn't be. Consider the Cyberball procedure we mentioned earlier, in which participants think they are being ignored and rejected by others in an online ball tossing game. When Cyberball is administered on computers in the lab, the experimenter can take time after the session to ensure the participant understands the exclusion was not real, and that any distress has been remedied before the participant leaves the laboratory. Indeed, research shows that this must be done very carefully. It is not enough to simply tell participants that the rejection was not real and bid them a good day. The Cyberball experience may have brought to mind similar examples from their past in which they were socially rejected, making them think that such experiences are more typical than they are (Ross, Lepper, & Hubbard, 1975). Undoing such effects requires a careful debriefing in which the possibility of such lingering effects is discussed (we discuss the postexperimental interview in more detail later in this chapter).

What about in online studies, where participants may drop out prior to debriefing, or quickly click-through to exit more quickly? In studies that pay participants by the minute, as many online platforms (e.g., Prolific, mTurk) do, there may even be an economic incentive for researchers to limit the time spent on debriefing. Cyberball participants may never come to learn that their exclusion was faked. Even if they do, it is very difficult to conduct the kind of detailed debriefing online that is necessary to ameliorate any negative effects. For these reasons, we suggest online data collection is generally inappropriate for experiments that require more thorough debriefing, either due to extensive deception or due to concerns about potential harm, unless it can be guaranteed that all participants who undergo the manipulation will be thoroughly debriefed (e.g., by conducting the experiment over teleconferencing platforms, such as Zoom, in which an experimenter is present for the entire session).

In sum, when designing experimental studies for use online, we caution that even high-quality judgment experiments should not be substituted for impact experiments simply because they are easier to implement. A prime example comes from the research on moral dilemmas described earlier. Although it would be far easier to ask online participants to respond to hypothetical scenarios (judgment experiment) than to decide whether to harm one mouse versus five (impact experiment), the answers derived from these studies answer distinctly different research questions: one about lay theories, the other about actual moral reasoning, respectively. Although it may be tempting to adapt designs in ways that facilitate online data collection, more important is whether the design adequately tests the hypothesis in question.

The Four Stages Of Experimentation

The process of planning an experiment consists of four basic stages: (1) setting the stage for the experiment, (2) constructing the independent variable, (3) measuring the dependent variable, and (4) planning the post-experimental follow-up. In this section, we will suggest ways of developing a sensible and practical modus operandi for each of those stages. We will focus on the impact experiment, which tends to be more complex and requires a wider scope of planning than the judgment experiment. Many of the lessons we convey, however, apply to any kind of experimental research.

Setting The Stage

In designing any experiment, a great deal of ingenuity and invention must be directed toward the context, or stage, for the manipulation of the independent variable. Because our participants tend to be intelligent, adult, curious humans, the setting must make sense to them. It not only must be consistent with the procedures for presenting the independent variables and measuring their impact but also can and should enhance that impact and help to justify the collection of the data.

Some experiments involve deception; if deception is used, the setting must include a sensible, internally consistent pretext or rationale for the research as well as a context that both supports and enhances the collection of the data and reduces the possibility of detection. This false rationale is often referred to as a cover story.

In a judgment experiment, the cover story is typically less elaborate and more straightforward than in an impact experiment. Although deception is sometimes used in a judgment experiment, it is usually minimal and aimed primarily at increasing the interest of the participants and providing a credible rationale for the data collection procedures and judgment task. As the authors predicted,

For example, Wages et al. (2020) tested the hypothesis that Black people are stereotyped as “reckless risk-takers” using the reverse-correlation method. Researchers generated images of faces, and participants selected the ones they thought looked the most responsible or reckless. The researchers then merged these faces to create prototypically responsible-looking and prototypically reckless-looking faces. Next, a new group of participants was invited to play an investment game. Participants learned about two “investors” and were asked to allocate money between them. One of the investors was depicted by the prototypically responsible-looking face, the other investor was depicted by the prototypically reckless-looking face. Because the faces were digital composites from the previous study they were of a degraded quality. What would be a reasonable cover story that would explain why participants were being asked to view such fuzzy images? The experimenters simply informed participants that they were images of previous participants who had played the role of investors and that the researchers had obscured the images to protect participants’ privacy. The cover story in this experiment was simple and straightforward and succeeded in providing a credible rationale for both the presentation of the stimulus and the collection of the data. (As predicted, participants allocated more money to the investor with the “responsible looking” face, which previous participants had perceived as more likely to be White than African American).

Providing a convincing rationale for the experiment is almost always essential, because participants attempt to make sense of the situation and to decipher the reasons for the experiment. A good cover story is one that embraces all the necessary aspects of the experiment in a plausible manner and thus eliminates speculation from a participant about what the experimenter really has in mind. It also should capture the attention of the participants so that they remain alert and responsive to the experimental events. This is not meant facetiously; if a cover story strikes the participants as being a trivial or silly reason for conducting an experiment, they may simply tune out. If the participants are not attending to the independent variable, it will have little impact on them.

The setting may be a relatively simple one, or it may involve an elaborate scenario, depending on the demands of the situation. Obviously, the experimenter should set the stage as simply as possible. If a simple setting succeeds in providing a plausible cover story and in capturing the attention of the participants, there is no need for greater elaboration. A more elaborate setting is sometimes necessary, especially in a high-impact experiment. For example, suppose researchers want to see if people will behave sadistically. They might achieve this goal by simply telling the participants they will have the opportunity to end the life of a few small insects. Yet the chances of observing actual sadistic behavior are enhanced if one has set the stage with a trifle more embellishment. This can be done by providing live fly larvae in individual cups, individually labeling the larvae with cute names, providing a realistic-looking coffee grinder that vibrates and makes grinding noises when activated, and leaving participants in privacy to place bugs in the grinder and flip a switch to grind them up, as Pfattcher et al. (2021) did in experiments on the effects of boredom on sadistic aggression.

The point we are making is that in a well-designed experiment, the cover story is an intricate and tightly woven tapestry. With this in mind, let us take another look at the Latané and Darley (1968) experiment. The major challenges presented by the hypothesis were to have an emergency occur in a believable and plausible manner and vary the number of bystanders, without introducing confounding factors. This was solved by (a) the cover story that participants would be discussing personal problems with other students; (b) having them do so over an intercom system while alone in a room; (c) prerecording the voices of the "other students," including, crucially, the one who had the seizure. This was done in a manner that was quite believable and made sense to participants. Indeed, when the seizure occurred, participants took it seriously; many gasped or exclaimed to themselves, "My God, he's having a fit" (p. 381). In other words, the cover story worked. The authors report that all but two participants believed the seizure was real.

The testing of some hypotheses is more difficult than others because of their very nature. But none is impossible; with sufficient patience and ingenuity a reasonable context can be constructed to integrate the independent and dependent variables regardless of the problems inherent in the hypothesis.

Constructing The Independent Variable

One of the most important and difficult parts of experimental design is constructing an independent variable that manipulates only what you want it to manipulate. The experimenter begins with what we will call the *conceptual variable,* which is a theoretically important variable that they think will have a causal effect on people's responses. In the Festinger and Carlsmith (1959) study discussed earlier, for example, the conceptual variable was cognitive dissonance caused by lying to another participant when external justification (the payment) was low. There are many ways to translate an abstract conceptual variable such as this into a concrete experimental operation; and indeed, dissonance has been manipulated in many ways. One of the most important parts of experimental design is to devise a procedure that "captures" the conceptual variable perfectly without influencing any other factors--which is often easier said than done. Indeed, critics expressed doubt that Festinger and Carlsmith had successfully manipulated cognitive dissonance, offering a number of alternative explanations, which led to further tests of those explanations (see Cooper, 2007 for a review). As a more recent example, there has been debate over the correct interpretation of facial feedback studies, which purport to show that people's facial expressions influence their emotional experience (Strack, Martin, & Stepper, 1988). Follow-up studies have shown not only that the effect is reliable, but have ruled out alternative explanations (Coles et al., 2022; Noah, Schul, & Mayo, 2018). These examples illustrate that single studies are rarely sufficient to confirm a hypothesis. As we noted earlier, it is important to show that multiple instantiations of the independent variable yield similar results.

The issue of standardization

Ideally, the empirical realization of an independent variable is forceful enough to have maximum impact and clear enough to generate the intended interpretation in all participants. To this end, experimental scripts and even videotaping the study can be enormously helpful for training experimenters on the intended manipulation. An added advantage, of course, is documenting study procedures; such records can be shared upon eventual publication and help other researchers assess and even replicate the original study procedures. With the advent of free and easy online storage, there is little reason for researchers not to make their materials publicly available; sharing data and materials is valuable to the scientific community and helps advance scientific progress.

There is, however, one crucial, yet frequently misunderstood, point: It is extremely important for all participants to be in the same psychological state as a result of the manipulation of the independent variable. This usually involves exposing all participants to the identical independent variable, with the assumption that they will interpret it in similar ways. There may be cases, however, in which some types of participants require a different "dose" to achieve the same psychological state--if this is well-documented and determined in advance of the study. A brief analogy to drug trials will make our point. When clinicians test the efficacy of an anti-anxiety medication, they administer a precise amount of the drug to ensure that each participant develops the same concentration of the drug in their bloodstream, an amount that may vary from participant to participant. Similarly, in social psychological experiments it may be necessary to "titrate" the independent variable in a similar manner. For example, Wilson et al. (2014) were interested in whether participants would administer themselves a painful electric shock during a 15-minute "thinking" period in which they were supposed to be entertaining themselves with their thoughts. All participants were administered a sample shock before the thinking period, so that they knew what they were "getting into" if they chose to shock themselves again. When pilot testing this procedure, the researchers found that women rated the sample shock to be more painful than men did. Therefore, the shocks given to women in the actual study were of slightly lower intensity than the shocks given to men, and as a result, women and men gave the sample shock similar pain ratings. That is, all participants began the thinking period with a similar understanding of how painful the shocks were. (As it happened, two-thirds of men and a quarter of women shocked themselves at least once during the thinking period.)

We anticipate that many experimenters will disagree with us, suggesting that standardization is the hallmark of an experiment. We agree, but exactly what is it that should be standardized? We suggest it should be what the participant understands, which might involve titration--which, of course, should be determined in advance of the study through pilot testing. A similar issue has come up in the context of the replication movement, namely what constitutes an "exact" replication of an earlier study. One answer is that the replication should follow the procedures of the earlier study as closely as possible, including how the independent variable was operationalized. But if the original manipulation was specific to a particular historical or cultural context, and the replication is being conducted in a different historical or cultural context, the matter is more complicated. For example, suppose that researchers today were interested in replicating the Festinger and Carlsmith (1959) study: Should they offer participants the same amounts of money ($.50 or $20) to lie to the next participant? Doing so would, in one sense, be faithful to the original study. But those amounts of money would mean something quite different to participants today than they did to participants in 1959, thus a case could be made to adjust the amounts for inflation. Doing so would maintain the psychological fidelity of the manipulation. We believe it makes more sense, when conducting a replication, to try to duplicate the psychological impact of an independent variable, rather than adhering strictly to a "do it exactly as they did" rule (Schwarz & Strack, 2014).

We return now to a discussion of independent variables and how they should be administered. Recall that the essence of an experiment is the random assignment of participants to experimental conditions. For this reason, it should be obvious that any characteristics that the participants bring to the experiment cannot be regarded as independent variables in the context of a true experiment. It is not infrequent to find an “experiment” purporting to assess the effects of a participant variable (such as level of self-esteem) on some behavior in a specific situation. It should be clear that although such a procedure may produce interesting results, it is not an experiment because the variable was not randomly assigned. Nonrandom assignment of participants to experimental conditions is not confined to the use of personality measures in lieu of experimental treatments. It can take place in more subtle ways. One of the most common occurs when the experimenter is forced to perform an “internal analysis” to make sense out of their data.

The term “internal analysis” refers to the following situation. Suppose that an experimenter has carried out a true experiment, randomly assigning participants to different treatment conditions. Unfortunately, the treatments do not produce any measurable differences on the dependent variable. In addition, suppose that the experimenter has had the foresight to include an independent measure of the effectiveness of the experimental treatment. Such manipulation checks are always useful in providing information about the extent to which the experimental treatment had its intended effect on each individual participant. Now, if the manipulation check shows no differences between experimental treatments, the experimenter may still hope to provide further tests of their hypothesis on an exploratory basis. That is, the manipulation check shows that for some reason the treatments were unsuccessful in creating the internal states in the participants that they were designed to produce. Since they were unsuccessful, one would not expect to see differences on the dependent variable. In this case, the experimenter may analyze the data based on the responses of the participants to the manipulation check, seeing whether participants who responded in the expected way to the manipulation check exhibited different results than those who did not. This is an internal analysis.

For example, Schachter (1959) attempted to alter the amount of anxiety experienced by his participants by varying the description of the task in which the participants were to engage. However, in some of the studies, many participants who had been given the treatment designed to produce low anxiety actually reported higher anxiety levels than some who had been given the treatment designed to produce high anxiety. From the results of an internal analysis of these data, it did seem that anxiety was related to the dependent variable. Again, these data can be useful and provocative, but because the effect was not due to the manipulated variable, no causal statement can be made. Although many of the highly anxious participants were made anxious by the high-anxiety manipulation, many were highly anxious on their own. Because people who become anxious easily may be different from those who do not, we are dealing with an individual difference variable. This means that we can no longer claim random assignment--and, in effect, we no longer have an experiment.

For this same reason, post-hoc exclusion of participants in experimental designs based on failed manipulation checks or attention checks is risky (Montgomery et al., 2018). On the one hand, it makes sense to drop people who failed to understand the instructions or clearly weren’t paying attention, to perform a fair test of one’s hypotheses. But people who fail such checks may be different from those who pass them; excluding their data would mean that the experimental condition is no longer the product of random assignment, but rather is contingent on third variables and individual differences. Thus, a safer approach is to include all participants. Doing so, we note, is similar to “intent to treat” analyses in clinical trials that include all participants in the analyses, even if they did not follow the instructions or complete all the measures. Doing so preserves random assignment.

Between- versus within-participant designs

Another decision facing the experimenter is whether to manipulate the independent variable on a between-participant or within-participant basis. In a between-participant design people are randomly assigned to different levels of the independent variable, as in the Latané and Darley study, in which participants believed that the number of students in the group discussion was 2, 3, or 6. In a within-participants design all participants receive all levels of the independent variable. For example, in studies examining whether the presence of close others reduces threat responses, participants in an fMRI chamber were asked to hold the hand of a romantic partner, a stranger, or no hand at all, in alternating blocks, while viewing a screen that indicated they were either safe or at risk (20%) of receiving an electric shock on their ankle. Participants exhibited lower neural threat responses when holding hands with their romantic partner, and this effect was stronger for couples with stronger relationships (e.g., Coan et al., 2006).

This leads us to another decision: how many participants to run. For instance, detecting that people who like eggs eat egg salad more often than people who don't like eggs--a seemingly obvious finding--requires approximately 50 participants per cell, or 100 participants total (Simmons et al., 2018). One straightforward way to increase statistical power (i.e., the odds of detecting an effect if it really exists) is to increase sample size; indeed, this can be a major advantage of online studies. Because bigger effects require fewer participants, if an effect size is known in advance, researchers can conduct a power analysis to determine how many participants are needed.

However, often the size of an effect is not known beforehand, or increasing the sample sizes is not feasible or desirable. In these cases, an alternative to larger sample sizes is choosing research designs that maximize the size of the effect (i.e., signal) and/or minimize noise. For this reason, within-participant designs are often preferred, because fewer participants are required to achieve sufficient statistical power. Imagine that Coan et al. (2006) had used a between-participants design, such that three separate groups of participants were randomly assigned to hold the hand of a romantic partner, a stranger, or no hand at all. The number of participants required to achieve a sufficient level of power would increase dramatically. For example, to detect a rather large effect size of .5 in a two-condition, between-participant design, with 80% power at alpha = .05, would require 128 participants (64 per condition). To detect the same effect size in a within-participant design, with identical power, would require 34 participants. One reason fewer participants are needed is because each participant serves as their own control; each person's responses in one condition are compared to that same person's responses in the other conditions. This controls for any number of individual difference variables that are treated as error variance in a between-participants design.

If a within-participant design is used it is important, of course, to vary the order of the experimental conditions, to make sure that the effects of the independent variable are not confounded with the order in which people receive the different manipulations. This is referred to as "counterbalancing," whereby participants are randomly assigned to get the manipulations in different orders. In the Coan et al. (2006) study, for example, the order of the blocks in which participants held the hand of their romantic partner, a stranger, or no hand was counterbalanced.

In many social psychological experiments within-participant designs are not feasible, because it would not make sense to participants to evaluate the same stimulus more than once under slightly different conditions. Obviously, Darley and Latané could not plausibly have participants witness the same emergency twice, once alone and again in the presence of bystanders. Nor could Festinger and Carlsmith tell the same participant that they would receive $1 and $20 to lie to the next participant. Thus, within-participants designs are preferable if at all possible, but in many studies--especially impact experiments--they are not feasible.

Avoiding participant awareness biases

It is arguably more challenging to perform a meaningful experiment in social psychology than in any other scientific discipline for one simple and powerful reason: In social psychology, we are testing our theories and hypotheses on adult human beings who are almost always intelligent, curious, and experienced. They are experienced in the sense that they have spent their entire lives in a social environment and--because of their intelligence and curiosity--they have formed their own theories and hypotheses about precisely the behaviors we are trying to investigate. That is to say, everyone in the world, including the participants in our experiments, is a social psychological theorist.

In a nutshell, the challenge (and the excitement) of doing experiments in social psychology lies in the quest to find a way to circumvent or neutralize the theories that the participants walk in with so that we can discover their true behavior under specifiable conditions, rather than being left to ponder behavior that reflects nothing more than how the participants think they should behave in a contrived attempt to confirm their own theory. One special form of participant awareness is closely related to the idea of "demand characteristics" as described by Orne (1962). The term refers to features introduced into a research setting by virtue of the fact that it *is* a research study and that the participants know that they are part of it. As aware participants, they are motivated to make sense of the experimental situation, to avoid negative evaluation from the experimenter, and perhaps even to cooperate in a way intended to help the experimenter confirm the research hypothesis (Sigall, Aronson, & Van Hoose, 1970). Such motivational states could make participants responsive to any cues--intended or unintended--in the research situation that suggest what they are supposed to do to appear normal or "to make the study come out right." It is for this reason that experimenters frequently employ deception, elaborate cover stories, and the like, in an attempt to keep participants unaware of the experimental manipulations in play.

Another aspect of the problem of demand characteristics and participant awareness is the possibility that the experimenter's own behavior provides inadvertent cues that influence the responses of the participants. In our experience novice researchers often dismiss this possibility; they smile knowingly and say, "Of course *I* wouldn't act in such a way to bias people's responses." Decades of research on expectancy effects, however, show that the transmission of expectations from researchers to participants is subtle and unintentional, and that this transmission can have dramatic effects on participants' behavior. It can occur even between a human experimenter and an animal participant; in one study, for example, rats learned a maze quickly when the experimenter thought they were good learners and slowly when the experimenter thought they were poor learners (Rosenthal, 1994; Rosenthal & Lawson, 1964).

Therefore, steps must be taken to avoid this transmission of the experimenter's hypotheses to the research participants. One way of doing so is to keep the experimenter unaware of the hypothesis of the research. The idea here is that if the experimenter does not know the hypothesis, they cannot transmit the hypothesis to the research participants. In our judgment, however, this technique is inadequate. One characteristic of good researchers--indeed of all intelligent humans--is that they are hypothesis-forming organisms. Thus, if not told the hypothesis, the research assistant, like a participant, attempts to discover one. Moreover, keeping the assistant in the dark reduces the value of the educational experience. Since many experimenters are undergraduates or graduate students, full participation is the most effective way of learning experimentation. Any technique involving the experimenter's ignorance of the hypothesis or a reduction in contact with the supervisor is a disservice to them. A more reasonable solution involves allowing the experimenters to know the true hypothesis but keeping them ignorant of the specific experimental condition of each participant. This is typically referred to as a "double-blind" study in which both the experimenter and participant are both unaware of the experimental condition. In theory, this is a simple and complete solution to the problem and should be employed whenever possible.

In a study by Wilson et al. (1993), for example, the independent variable was whether people were asked to think about why they felt the way they did about some art posters, to examine the effects of introspection on attitude change and satisfaction with consumer choices. Participants were told that the purpose of the study was to examine the different types of visual effects that people like in pictures and drawings and that they would be asked to evaluate some posters. The critical manipulation was whether people were randomly assigned to write why they felt the way they did about each poster (the reasons condition) or why they had chosen their major (the control condition). To assign people to condition randomly, the experimenter simply gave them a questionnaire from a randomly ordered stack. To make sure the experimenter did not know whether it was the reasons or control questionnaire, an opaque cover sheet was stapled to each one. The experimenter left the room while the participant completed the questionnaire, and thus throughout the experiment was unaware whether the participant was in the reasons or control condition. Today, when many laboratory studies incorporate computers as part of the data collection process, double blinding can be easily accomplished by having the computer administer the key manipulation (randomly assigned). This comes with the risk, however, that participants will gloss over the key text that delivers the manipulation, reducing its effectiveness. For complex manipulations, it is often better to have an experimenter deliver them verbally.

In such cases, including ones in which the experimental manipulations cannot be delivered simply by having people read written instructions, it is more difficult to keep the experimenter unaware of condition. In studies on temporal cognition, for example, the critical manipulation is whether people are thinking about an event from the perspective of the past, present, or future. This could be conveyed in written form, but there is a risk that participants will not read the instructions carefully enough, missing the crucial information about the temporal distance. One solution to this problem is to make an audio or videorecording of the instructions, and to keep the experimenter unaware of which recorded instructions each participant receives (e.g., Bruehlman-Senecal & Ayduk, 2015).

In other studies, however--particularly high impact ones--the experimenter must deliver the independent variable in person, making it more difficult for them to be unaware of participant's experimental condition. For example, in the Furrer et al. (in press) studies of the illusion of unfairness mentioned earlier, the independent variable was which of two participants got to flip a coin to determine who got a good versus bad outcome. In studies such as these, where it is necessary for the experimenter to "deliver" the independent variable, several steps can still be taken to avoid demand characteristics, participant awareness biases, and experimenter expectancy effects. First, the experimenter should be kept ignorant of people's condition until the precise moment the manipulation is delivered. That is, in most studies, the experimenter need not know what condition the participant is in until the crucial manipulation occurs. In one of the Furrer et al. (in press) studies, for example, the experimenter explained the study in detail to the two participants without knowing which one would flip the coin. At the point at which the coin flip was to occur, the experimenter randomly assigned one of them to be the flipper.

This is only a partial solution because the experimenter loses their ignorance midway through the experiment. However, if the experimenter left the room immediately after the recitation and a different experimenter (unaware of the participant's experimental condition) collected the data, this solution would approach completeness. The use of multiple experimenters, each ignorant of some part of the experiment, offers a solution that is frequently viable. In the case of the Furrer et al. (in press) study, the experimenter had minimal contact with the participants after assigning them to condition; as soon as the coin was flipped, participants completed the dependent measures by themselves on a computer. Nevertheless, to make sure that even this minimal contact did not introduce experimental bias, the researchers conducted versions of the study online in which all the instructions and condition assignments were delivered on a computer (as discussed earlier, similar results were found in the laboratory and the online studies).

Returning to the more general issue of demand characteristics, it should be clear that the most effective type of deception in an impact experiment involves the creation of an independent variable as an event that appears not to be part of the experiment at all. Creating such an independent variable not only guarantees that the participant will not try to interpret the researcher's intention but also that the manipulation has an impact on the participant. Several classes of techniques have been used successfully to present the independent variable as an event unrelated to the experiment. Perhaps the most effective is the "accident" or "whoops" manipulation, in which the independent variable is presented as part of what appears to be an accident or unforeseen circumstance. Wilson, Hodges, and LaFleur (1995) used a variation on this procedure to influence people's memory for behaviors performed by a target person. These researchers showed people a list of positive and negative behaviors the target person had performed and then wanted to make sure that people found it easiest to remember either the positive or negative behaviors. They did so by simply showing people either the positive or negative behaviors a second time. The danger of this procedure, however, is that it would be obvious to people that the researchers were trying to influence their memory. If the experimenter had said, "OK, now we are going to show you only the positive (negative) behaviors again," participants would undoubtedly have wondered why and possibly figured out that the point was to influence their memory for these behaviors. To avoid this problem, the experimenter told people that they would see all the behaviors again on slides. After only positive (or negative) ones had been shown, it just so happened that the slide projector malfunctioned. The projector suddenly went dark, and after examining it with some frustration, the experimenter declared that the bulb was burned out. He searched for another for a while, unsuccessfully, and then told participants that they would have to go on with the study without seeing the rest of the slides. By staging this "accident," the researchers ensured that people were not suspicious about why they saw only positive or negative behaviors a second time.

Another way in which researchers can inadvertently bias results comes later in the process, after the experiment is in the analysis stage. Researchers who expect a certain outcome may unconsciously make decisions during data analysis--such as which participants to include, or what variables to analyze--which favor their hypothesis (e.g., confirmation bias). One solution to this is preregistration, which is registering one's hypothesis and/or analysis plan before the study is conducted (Nosek et al., 2018). Preregistration can be very helpful in cases where the "correct" analysis strategy is unclear (or does not exist; Silberzahn et al., 2018) or when making unusual analytic decisions, such as planned one-tailed tests or predicting a three-way interaction. They can also be a helpful communication tool for research teams, such as in adversarial collaborations where researchers with competing hypotheses come together to conduct a joint study (e.g., Killingsworth, Kahneman, & Mellers, 2023). However, while preregistrations can increase transparency, they are often time-consuming and it can be difficult to determine in advance what the “correct” analysis is. Audits show most researchers deviate from their preregistrations (Claesen et al., 2021), and the purpose of preregistration itself is still evolving (e.g., Ledgerwood, 2018). Indeed, some argue that exploratory data analyses have greater value than preregistered ones (Rubin & Donkin, 2022).

Pilot testing the impact of the independent variable

One of the most frequently misunderstood aspects of experimentation is the amount of pretesting that is often required to make sure that the independent variable is having the desired effect. Researchers do the best they can to devise an independent variable that manipulates what they intend to manipulate, but often the manipulation falls flat: Participants don’t understand it or interpret in a way that the researchers did not anticipate. This is why it is crucial to pilot test an independent variable. For example, in the Wilson et al. (1995) study mentioned earlier, in which the researchers staged a malfunction of a slide projector, a good deal of pretesting was required to “fine tune” this manipulation. Different versions of the manipulation were tried before one was found that had the desired effect.

It is important to be clear what we mean here by the “desired effect.” Indeed, it can mean two things: (a) whether the researchers manipulated what they intended to manipulate and (b) whether the independent variable had the predicted effect on the dependent variable. In the first case, an experiment cannot test a hypothesis unless the independent variable manipulates what it is supposed to manipulate. And, equally important, that it manipulates that and only that. For example, in the Wilson et al. (1995) study, the point was to see what happens when people analyze the reasons for their impressions of a person and either positive or negative thoughts about that person are most accessible in memory. The hypotheses of the study could only be tested if the manipulation of people’s memory succeeded in making positive or negative thoughts more accessible. The ability to play with a design so that the manipulations change the right variables is a skill similar to that of a talented director who knows exactly how to alter the staging of a play to maximize its impact on the audience.

It is another matter, however, if the manipulation works as intended but does not influence the dependent variable in the predicted manner. This can happen simply because the researcher’s hypothesis is wrong. The manipulation might work exactly as intended (as indicated on a manipulation check), but have a different effect on the dependent variable than predicted. This *is* informative because it suggests that the hypothesis might be wrong. The catch is that it can be difficult to tell whether an experiment is not working because the manipulation is ineffective or because of a faulty hypothesis. The answer to this question often becomes clear only after extensive tinkering and restaging of the experimental situation through pilot testing. Such testing should include not only checking that the manipulation works as intended, but also that it does so cleanly--that is, by measuring and checking for variables that the manipulation should *not* affect (such as teacher attention in our earlier breakfast-eating example).

Ambiguity over pilot testing, however, has been one of the concerns of the open science movement (e.g., Albers & Lakens, 2018). One issue is that researchers might engage in extensive pilot testing of different versions of their independent variable, and report only the one that "works," in the sense of our second meaning of "desired effects" above. This would increase the likelihood of Type 1 errors, whereby researchers stumble on a positive effect by chance, when in fact their hypothesis is false (Simonsohn, Nelson, & Simmons, 2014). Further, if researchers don't report the pilot testing in their published reports, readers will not have the benefit of learning about the false starts, i.e., the versions of the independent variable that did not work.

On the other hand, validating experimental manipulations is essential, and researchers who fail to pilot test their independent variables do so at their own peril. Ellsworth (2010) stated this well:

> Moving straight to the real experiment without pilot testing can be even more costly. I've read MA and PhD theses in which the treatment didn't work: the negative mood induction made Subjects laugh because it was too corny; the "popular" CDs that Subjects were given to choose from were 10 years out of date, because they were the exact same ones some researcher used 10 years ago and the Subjects never heard of any of them and had no preference; nobody believed your story that the IQ test they were taking was really race neutral. Finding out the answers to questions like these is what pilot testing is all about. It means that when you finally do run the actual study, the treatments mean what you meant them to mean, the measures measure what, they are supposed to . . . Like all fine craftsmanship, it takes skill, work, and time; but the effort is worth it because it saves the researcher from running an entire study that gets null results because of flaws that could have been discovered and corrected in advance. (pp. 90-91)

Our advice is to engage in extensive pilot testing, for the reasons Ellsworth relates, especially for complex independent variables or ones that are difficult to instantiate. Now that it is standard to include online supplementary materials accompanying a published article, we likewise suggest that researchers report the pilot testing they have conducted. Doing so not only provides additional validation for the experimental manipulation in question, but gives readers the benefit of learning about the false starts, i.e., the versions of the independent variable that did not work.

Once it becomes clear (from pilot testing) that the manipulation is working as intended but the hypothesis is off the mark, a second talent comes into play: The ability to learn from one's mistakes. Some of the most famous findings in social psychology did not come from reading the literature and deducing new hypotheses, or from "aha" insights while taking a shower. Rather, they came about from the discovery that one's hypotheses were wrong, and that the data suggest a very different hypothesis--one that is quite interesting and worth pursuing. For this reason, it is crucial to analyze one's data thoroughly, even if such analyses were not preregistered. The data might reveal a hypothesis that is more interesting than the one that inspired the study.

Choosing the number of independent variables

We have been talking thus far about the independent variable in social psychological experiments as if it were a simple two-level variation on a single dimension. Yet many, if not most, experiments involve procedures that simultaneously manipulate two or more variables. Once one has taken the time and trouble of setting up a laboratory experiment, recruiting participants, and training research assistants, it seems only efficient to take the occasion to assess the effects of more than one experimental treatment.

There are no pat answers to the question of how many independent variables can or should be manipulated at one time, but our own rule is that an experiment should be only as complex as is required for important relationships to emerge in an interpretable manner. Sometimes it is essential to vary more than one factor because the phenomenon of interest appears in the form of an interaction. Petty, Cacioppo, and Goldman (1981), for example, hypothesized that the way in which people process information in a persuasive communication depends on the personal relevance of the topic. When the topic was highly relevant, people were predicted to be most influenced by the strength of the arguments in the communication, whereas when it was low in relevance, people were predicted to be most influenced by the expertise of the source of the communication. To test this hypothesis the authors had to manipulate (a) the personal relevance of the topic, (b) the strength of the arguments in the message, and (c) the expertise of the source of the message. Only by including each of these independent variables could the authors test their hypothesis, which was confirmed in the form of a three-way interaction. Fortunately, although increasing the number of independent variables itself increases the required sample sizes, interaction effects in experimental designs where all independent variables are randomly assigned (i.e., measured without error) do not necessarily require greater sample sizes to detect than main effects. In correlational designs, where predictors are measured, such measurement inevitably introduces measurement error. When those predictors are multiplied together to produce an interaction term, the resulting interaction term also multiplies the measurement error. This vastly reduces power to detect interactions. However, in experimental designs, because there is no measurement error in randomly assigned predictors, multiplying experimental conditions to produce an interaction term does not result in any additional measurement error. As a result, unlike correlational designs that use measured predictors, experiments do not lead to a decrease in statistical power to detect interaction effects. This can be seen in simulations that calculate power to detect interaction effects in correlational designs (i.e., where predictors are measured with error) versus experimental designs (i.e., where predictors consist of known groups, and are thus measured without error).

An exception, of course, is examining interactions between randomly-assigned variables and measured variables (e.g., experimental condition moderated by an individual difference variable), or when there are a priori theoretical reasons to believe that the size of a specific interaction effect should be smaller than the respective main effects. For example, imagine a lab study in which participants are asked to walk to a different building and the research assistant offers them an umbrella. The umbrella, incidentally, has either their own university's logo on it or a rival's. We might expect two large main effects: participants strongly prefer the umbrella with their university's logo, and they strongly prefer an umbrella more when it's raining than when it's sunny. We might also expect a smaller interaction, namely that the preference for their university logo is less strong when it's raining (that is, when it's raining they're more willing to accept the umbrella with the rival school's logo). In this case, we would need more power to detect the small interaction effect than we would to detect the large main effects.

Measuring The Dependent Variable

What is the best way to measure the hypothesized effects of the independent variable? It is possible to imagine a continuum ranging from behaviors of great importance and consequence for the participant down to the most trivial paper-and-pencil measures about which the participant has no interest. At one extreme the experimenter could measure the extent to which participants perform a great deal of tedious labor for a fellow student (as a reflection of, say, their liking for that student). At the other extreme one could ask them to circle a number on a scale entitled, "How much did you like that other person who participated in the experiment?" Close to the behavioral end of the continuum would be a measure of the participant's commitment to perform a particular action without actually performing it. This is typically called a "behavioroid" measure.

Many social psychological studies are about social behavior: how people treat each other and how they respond to the social world. The goal is not to explain and predict which number people will circle on a scale or which button on a computer they will press, but people's actual behavior toward another person or the environment. In such cases, the first choice of a dependent measure in a social psychological experiment is overt behavior. The ideal measure of helping behavior is, well, whether people actually try to help a person in need. The ideal measure of discrimination is the way in which members of different groups treat each other, the ideal measure of attitude change is behavior toward an attitude object, and the ideal measure of interpersonal attraction is affiliative behaviors between two individuals. If you pick up a copy of a recent social psychological journal, however, you will find that measures of actual behavior are hard to come by (Baumeister, Vohs, & Funder, 2007; de la Haye, 1991). The dependent measures are more likely to be such things as questionnaire ratings of people's thoughts, attitudes, emotions, and moods; their recall of past events; or the speed with which they can respond to various types of questions.

There are four main reasons why social psychologists often measure things other than actual behavior. The first is convenience: It is much easier to give people a questionnaire on which they indicate how much they like a target person, for example, than to observe and code their actual behavior toward that person. Further, it is far more convenient to conduct studies online instead of in person, where it is impossible to measure people's behavior (other than their keystrokes on a computer or mobile device). Of course, convenience is no excuse for doing poor science, and the assumption that questionnaire responses are good proxies for actual behavior should not be taken on faith. In the early years of attitude research, for example, researchers assumed that people's questionnaire ratings of their attitudes were good indicators of how they would actually behavior toward the attitude object. But it became apparent that this was often not the case (e.g., Wicker, 1969), and many researchers devoted their energies to discovering when questionnaire measures of attitudes predict behavior. A large literature on attitude-behavior consistency was the result, and it is now clear that self-reported attitudes predict behavior quite well under some circumstances but not others (e.g., Fazio, 1990; Wilson, Dunn, Kraft, & Lisle, 1989).

Needless to say, there are some situations in which obtaining a direct measure of the behavior of interest is not simply inconvenient, it is virtually impossible--for example, if the study is being conducted online. Another reason is if the researchers are studying highly personal, private behaviors, such as whether people practice safer sex. Because this is impossible to observe directly, researchers have used self-reports as proxy, or measured behavior that was a good indication of people's intentions. At the close of an experiment on safer sex, for example, the experimenter, while leaving the room, indicated that the participants, if they wanted, could purchase condoms (at a bargain price) by helping themselves from huge a pile of condoms on the table and leaving the appropriate sum of money (Aronson, Fried, & Stone, 1991). Although the participants had no way of suspecting that their behavior was being monitored, as soon as they left the laboratory, the experimenter returned and re-counted the condoms on the table to ascertain exactly how many they had purchased. Admittedly, the number of condoms purchased is not quite as direct a measure as the actual use of condoms, but especially given the fact that this measure was consistent with self-report measures, it seems like a reasonable proxy.

An additional reason why nonbehavioral measures are often used is that, in many situations, they can be a more precise measure of intervening processes than overt behavior. Behavior is often complex and multidetermined, making it difficult to know the exact psychological processes that produced it. For example, suppose in an experiment an accomplice (posing as a fellow participant) either praises the participant, implying that they are brilliant, or insults the participant, implying that they are less than brilliant. Suppose our dependent variable is how much the participant likes the accomplice. We can measure it by handing participants a rating scale and asking them to rate their liking for the accomplice, from +5 to —5. Or, on a more behavioral level, we can observe the extent to which the participant makes an effort to join a group to which the accomplice belongs. This latter behavior seems to reflect liking, but it may reflect other things instead. For example, it may be that some participants in the "insult" condition want to join the group to prove to the accomplice that they are brilliant. Or it may be that some want an opportunity to see the insulting person again so that they can return the favor. Neither of these behaviors reflects liking, and consequently, may produce results different from those produced by the questionnaire measure.

Lastly, some psychological states are best measured by self-report instruments--that is, by asking people--than by observing their behavior. In recent years, for example, there has been a considerable amount of interest in human happiness, such as what causes it, how well people can predict it, and whether it can be changed (e.g., Diener & Biswas-Diener, 2008; Gilbert, 2006; Wilson & Gilbert, 2003). Researchers have conducted a great deal of psychometric work on how best to measure how happy people are, and it turns out that the most valid and reliable way is to ask them (Andrews & Robinson, 1991; Diener, 1994; Fordyce, 1988). Indeed, research shows that happiness and other specific emotions cannot be distinguished by physiological signatures (Siegel et al., 2018), facial expressions (Gendron et al., 2014; Barrett et al., 2019), or neural activity (Lindquist et al., 2015). Thus, in some cases self-report instruments are the best measure of the phenomenon researchers are trying to assess.

Nonetheless, it is important to note some limitations of questionnaire measures. Most fundamentally, people may not know the answer to the questions they are asked. This is especially true of "why" questions, whereby people are asked to report the reasons for their behavior and attitudes. Rather than reporting accurately, people might be relying on cultural or idiosyncratic theories about the causes of their responses that are not always correct (Nisbett & Wilson, 1977; Wilson, 2002).

Disguising the measure

Even if people know the answer to a question, they may not answer truthfully. As previously mentioned, people might distort their responses due to self-presentational concerns or because they have figured out the hypothesis and want to tell the experimenters what they want to hear. It is thus often important to disguise the measurement of the dependent variable. This presents problems very similar to those involved in attempting to disguise the independent variable, as discussed in the earlier section on guarding against demand characteristics. Again, there are several classes of solutions that can be applied to the problem of disguising the dependent variable. One approach is to measure the dependent variable in a setting that participants believe is totally removed from the remainder of the experiment. For example, in research on intrinsic motivation it is common to assess people's interest in an activity by observing how much time they spend on that activity during a "free time" period. Participants believe that this time period is not part of the experiment and do not know that they are being observed. Lepper et al., (1973), for instance, measured children's interest in a set of felt-tip pens by unobtrusively observing how much time they spent playing with the pens during a free-play period of their preschool class.

Another approach is to tell participants that the dependent variable has nothing to do with the purpose of the study. For example, Hsee and Ruan (2016) were interested in whether people would voluntarily experience pain by clicking on prank pens that delivered mild electric shocks. In the certain condition, participants knew which of 10 pens would deliver shocks and which would not, whereas in the uncertain condition, they did not know which pens would deliver shocks. As predicted, participants were more likely to click on pens in the uncertain condition, presumably out of curiosity as to which ones would shock them. But maybe participants figured out that that was the hypothesis and clicked on more pens to be "good participants." To rule this out, the researchers told participants that the purpose of the study had nothing to do with the pens, which were supposedly left over from a prior study. Instead, participants thought they would be evaluating another set of stimuli that had not yet arrived. The experimenter told participants that while waiting they could kill time by playing with the pens. Because participants didn't think that the pens were part of the study, and because they thought they were unobserved (the experimenter sat on the other side of the room and then exited to get the "real" stimuli), we can be reasonably sure that their decision of whether to shock themselves was not determined by demand characteristics. A similar approach is to administer the dependent variable after the study is supposedly over, so that participants are unaware that it is the variable of interest (e.g., Aronson et al., 1991).

The Postexperimental Follow-Up

The experiment does not end when the data have been collected. Rather, the prudent experimenter will want to remain with the participants to talk and listen in order to accomplish four important goals: (a) to ensure that the participants are in a good and healthy frame of mind; (b) to be certain that the participants understand the experimental procedures, the hypotheses, and their own performance so that they gain a valuable educational experience as a result of having participated; (c) to avail themselves of the participant's unique skill as a valuable consultant in the research enterprise; that is, only the participants know for certain whether the instructions were clear, whether the independent variable had the intended impact on them, and so on; (d) to probe for any suspicion on the part of the participants, such as whether they believed the cover story.

It is impossible to overstate the importance of the postexperimental interview, also known as debriefing. The experimenter should never conduct it in a casual or cavalier manner. Rather, the experimenter should probe gently and sensitively to be certain that all of the above goals are accomplished. This is especially and most obviously true if any deception has been employed. In this case, the experimenter needs to learn if the deception was effective or if the participant was suspicious in a way that could invalidate the data based on their performance in the experiment. Even more important, where deception was used, the experimenter must reveal the true nature of the experiment and the reasons why deception was necessary. Again, this cannot be done lightly. People do not enjoy learning that they have behaved in a naive or gullible manner. The experimenter not only must be sensitive to the feelings and dignity of the participants but also should communicate this care and concern to them. Then, in explaining why the deception was necessary, the experimenter not only is sharing their dilemma as an earnest researcher (who is seeking the truth through the use of deception) but also is contributing to the participants' educational experience by exploring the process as well as the content of social psychological experimentation.

Although it is important to provide people with a complete understanding of the experimental procedures, this is not the best way to begin the postexperimental session. To maximize the value of the participants as consultants, it is first necessary to explore with each the impact of the experimental events. The value of this sequence should be obvious. If we tell the participants what we expected to happen before finding out what the participants experienced, they may be reluctant to report that they thought the procedures were pallid, misguided, or worthless. Moreover, if deception was used, the experimenter, before revealing the deception, should ascertain whether the participant was suspicious and whether particular suspicions were of such a nature as to invalidate the results.

It is best to explore the feelings and experiences of the participants in a gentle and gradual manner (Bargh & Chartrand, 2000). Why the need for gradualness? Why not simply ask people if they suspected that they were the victims of a hoax? Participants may not be responsive to an abrupt procedure for a variety of reasons. First, if a given person *did* see through the experiment, they may be reluctant to admit it out of a misplaced desire to be helpful to the experimenter. Second, as mentioned previously, because most of us do not feel good about appearing gullible, some participants may be reluctant to admit that they can be easily fooled. Consequently, if participants are told pointedly about the deception, they might imply that they suspected it all along, to save face. Thus, such an abrupt procedure may falsely inflate the number of suspicious participants and may, consequently, lead the experimenter to abandon a perfectly viable procedure. Moreover, as mentioned previously, abruptly telling people that they have been deceived is a harsh technique that can add unnecessarily to their discomfort and, therefore, should be avoided.

The best way to begin a postexperimental interview is to ask the participants if they have any questions. If they do not, the experimenter should ask if the entire experiment was perfectly clear--the purpose of the experiment as well as each aspect of the procedure. The participants should then be told that people react to things in different ways, and it would be helpful if they would comment on how the experiment affected them, why they responded as they did, and how they felt at the time, for example. Then each participant should be asked specifically whether there was any aspect of the procedure that they found odd, confusing, or disturbing.

By this time, if deception has been used and any participants have any suspicions, they are almost certain to have revealed them. Moreover, the experimenter should have discovered whether the participants misunderstood the instructions or whether any responded erroneously. If no suspicions have been voiced, the experimenter should continue: “Do you think there may have been more to the experiment than meets the eye?” This question is virtually a giveaway. Even if the participants had not previously suspected anything, some will probably begin to suspect that the experimenter was concealing something. In our experience, we have found that many people will take this opportunity to say that they did feel that the experiment, as described, appeared too simple (or something of that order). This is desirable; whether the participants were deeply suspicious or not, the question allows them an opportunity to indicate that they are not the kind of person who is easily fooled. The experimenter should then explore the nature of the suspicion and how it may have affected the participant’s behavior. From the participant’s answers to this question, the experimenter can make a judgment as to how close a participant’s suspicions were to the actual purpose of the experiment and, consequently, whether the data are admissible. Obviously, the criteria for inclusion should be both rigorous and rigid and should be set down before the experiment begins; the decision should be made without knowledge of the participants’ responses on the dependent variable.

The experimenter should then continue with the debriefing process by saying something like this: "You are on the right track, we *were* interested in exploring some issues that we didn't discuss with you in advance. One of our major concerns in this study is …" The experimenter should then describe the problem under investigation, specifying why it is important and explaining clearly exactly how the deception took place and why it was necessary. Again, experimenters should be generous in sharing their own discomfort with the participant. They should make absolutely certain that the participant fully understands these factors before the postexperimental session is terminated.

It is often useful to enlist the participant's aid in improving the experiment. Often the participant can provide valuable hints regarding where the weaknesses in the manipulation occurred and which one of these caused competing reactions to the one the experimenter intended. These interviews can and should, of course, be continued during the time the experiment is being run, but it is usually during pretesting that the most valuable information is obtained.

Finally, regardless of whether deception is used, the experimenter must attempt to convince the participants not to discuss the experiment with other people until it is completed. This is a serious problem because even a few sophisticated participants can invalidate an experiment. Moreover, it is not a simple matter to swear participants to secrecy; some have friends who may subsequently volunteer for the experiment and who are almost certain to press them for information. Perhaps the best way to reduce such communication is to describe graphically the colossal waste of effort that would result from experimenting with people who have foreknowledge about the procedure or hypothesis of the experiment and who thus can rehearse their responses in advance. The experimenter should also explain the damage that can be done to the scientific enterprise by including data from such participants. It often helps to provide participants with some easy but unrevealing answers for their friends who ask about the study (e.g., "it was about social perception"). If we experimenters are sincere and honest in our dealings with the participants during the postexperimental session, we can be reasonably confident that few will break faith.

To check on the efficacy of this procedure, Aronson (1966) enlisted the aid of three undergraduates who each approached three acquaintances who had recently participated in one of his experiments. The accomplices explained that they had signed up for that experiment, had noticed the friend's name on the sign-up sheet, and wondered what the experiment was all about. The experimenter had previously assured these accomplices that their friends would remain anonymous. The results were encouraging. Despite considerable urging and cajoling on the part of the accomplices, none of the former participants revealed the true purpose of the experiment; two of them went as far as providing the accomplices with a replay of the cover story, but nothing else.

What if the participant *has* been forewarned before entering the experimental room? That is, suppose a participant does find out about the experiment from a friend who participated previously. Chances are the participant will not volunteer this information to the experimenter before the experiment. Once again, we as experimenters must appeal to the cooperativeness of the participant, emphasizing how much the experiment will be compromised if people knew about it in advance. We cannot overemphasize the importance of this procedure as a safeguard against the artifactual confirmation of an erroneous hypothesis because of the misplaced cooperativeness of the participant. If the participants are indeed cooperative, they will undoubtedly cooperate with the experimenter in this regard also and will respond to a direct plea of the sort described.
Experimental social psychologists have been deeply concerned about the ethics of experimentation for a great many years precisely because our field is constructed on an ethical dilemma. Basically,

the dilemma is formed by a conflict between two sets of values to which most social psychologists subscribe: a belief in the value of free scientific inquiry and a belief in the dignity of humans and their right to privacy. We will not dwell on the historical antecedents of these values or on the philosophical intricacies of the ethical dilemma posed by the conflict of these values. It suffices to say that the dilemma is a real one and cannot be dismissed either by making pious statements about the importance of not violating a person's feelings of dignity or by glibly pledging allegiance to the cause of science. It is a problem every social psychologist must face squarely, not just once, but each time they construct and conduct an experiment, since it is impossible to delineate a specific set of rules and regulations governing all experiments. In each instance the researcher must decide on a course of action after considering the importance of the experiment and the extent of the potential injury to the dignity of the participants.

Obviously, some experimental techniques present more problems than others. In general, experiments that employ deception cause concern because lying, in and of itself, is problematic. Similarly, procedures that cause pain, embarrassment, guilt, or other intense feelings present obvious ethical problems. In addition, any procedure that enables the participants to confront some aspect of themselves that may not be pleasant or positive is of deep ethical concern. For example, participants in Asch's (1951) classic study learned that they could conform in the face of implicit group pressure. Many of Milgram's (1974) participants learned that they could be pressured to obey an authority even when such obedience involved (apparently) inflicting severe pain on another human being. Some of Darley and Latané's (1968) participants learned they would not always help in an emergency.

It can be argued that such procedures are therapeutic or educational for the participants. For example, Darley and Latané (1968) reported that during the debriefing, all their participants "felt they understood what the experiment was about and indicated that they thought the deceptions were necessary and justified. All but one felt they were better informed about the nature of psychological research in general" (p. 382). But this does not, in and of itself, justify the procedure primarily because the experimenter could not possibly know in advance that it would be therapeutic for all participants. Moreover, it is arrogant for the scientist to decide that they will provide people with a therapeutic experience without their explicit permission.

The use of deception, when combined with the possibility of "self-discovery," presents the experimenter with a special kind of ethical problem. In a deception experiment it is impossible, by definition, to attain informed consent from the participants in advance of the experiment. For example, how could Darley and Latané attain informed consent from their participants without revealing aspects of the procedure that would have invalidated any results they obtained? An experimenter cannot even reveal in advance that the purpose of an experiment is the study of conformity or obedience without influencing the participant to behave in ways that are no longer "pure." Moreover, we doubt that the experimenter can reveal that deception might be used without triggering vigilance and, therefore, adulterating the participant's response to the independent variable.

It is worth noting that there have been some empirical investigations of the impact of deception experiments on participants. These studies have generally found that people do not object to the kinds of mild discomfort and deceptions typically used in social psychological research (e.g., Christensen, 1988; Sharpe, Adair, & Roese, 1992; Smith & Richardson, 1983). If mild deception is used, and time is spent after the study discussing the deception with participants and explaining why it was necessary, the evidence is that people will not be adversely affected. Nonetheless, the decision as to whether to use deception in a study should not be taken lightly, and alternative procedures should always be considered.

As we noted in our discussion of the postexperimental interview, it is critical to explain to participants in a deception study, at its conclusion, the true nature of the procedures and the reasons for them. We strongly recommend, however, that a thorough explanation of the experiment be provided regardless of whether deception or stressful procedures are involved. The major reason for this recommendation is that we cannot always predict the impact of a procedure; occasionally, even procedures that appear to be completely benign can have a powerful impact on some participants. An interesting example of such an unexpectedly powerful negative impact comes from a series of experiments on social dilemmas by Dawes and his students (Dawes, McTavish, & Shaklee, 1977). In these experiments, typically, the participant must decide between cooperating with several other people or "defecting." The contingencies are such that if all participants choose to cooperate, they all profit financially; however, if one or more defects, defection has a high payoff and cooperation produces little payoff. Each person's response is anonymous and remains so. The nature of the decision and its consequences is fully explained to the participants at the outset of the experiment. No deception is involved.

Twenty-four hours after one experimental session, a participant (who had been the sole defector in his group and had won $19) telephoned the experimenter trying to return his winnings so that it could be divided among the other participants (who, because they chose to cooperate, had each earned only $1). In the course of the conversation, he revealed that he felt miserable about his greedy behavior and that he had not slept all night. After a similar experiment, a woman who had cooperated while others defected revealed that she felt terribly gullible and had learned that people were not as trustworthy as she had thought. To alleviate this kind of stress, Dawes went on to develop an elaborate and sensitive follow-up procedure.

We repeat that these experiments were selected for discussion precisely because their important and powerful impact could not have been easily anticipated. We are intentionally not focusing on experiments that present clear and obvious problems like the well-known obedience study (Milgram, 1974). We have purposely selected an experiment that involves no deception and is well within the bounds of ethical codes. Our point is simple but important. No code of ethics can anticipate all problems, especially those created through participants discovering something unpleasant about themselves or others during an experiment. However, we believe a sensitive postexperimental interview conducted by a sincere and caring experimenter not only instructs and informs, but also provides important insights and helps reduce feelings of guilt or discomfort generated by such self-discovery (see Holmes, 1976a, 1976b; Ross, Lepper, & Hubbard, 1975).

CONCLUDING COMMENTS

Previous versions of this chapter discussed the virtues of laboratory experiments and how to conduct them. In this version, we have expanded that discussion to include experiments conducted online, which have become increasingly common. Indeed, several authors have noted—and decried—a decrease in the frequency with which social psychologists conduct laboratory experiments (Anderson et al., 2019; Ellsworth, 2010). It is worth noting the reasons for this, and offering some suggestions.

One reason, as we noted earlier, is that there are advantages to conducting experiments online. For the time and expense of one laboratory experiment, a graduate student could run many online studies, often with considerably larger and more diverse samples. A potential side effect of this, of course, is an increase in judgment experiments (which are easily adapted to online administration) over impact experiments, which take greater creativity to accomplish in online settings. Indeed, some impact studies cannot or should not be conducted online. We hope that researchers will carefully consider this trade off, and recognize that often there is often no substitute for the carefully controlled laboratory experiment.

But there are other barriers to conducting such studies that are worth mentioning. One is overreach by IRBs. On the one hand, difficult ethical issues can be involved in psychological research, such as the example we gave earlier of using the Cyberball procedure in online studies. It is necessary and useful to have checks and balances to prevent ethical lapses, and IRBs often serve that purpose well. However, many researchers are frustrated by the fact that they must submit time-consuming paperwork to conduct even the most innocuous studies. Indeed, research suggests that as incidents of unethical research become rarer, IRBs may broaden their criteria of what constitutes problematic research (Levari et al., 2018), objecting to things that have little to do with the welfare of the participants. Ellsworth (2010) gives the example of a researcher at a U.S. university who proposed to do a study in Japan. The IRB at their university denied permission to do the study, because participants were to be given a local Japanese phone number to call if they had any concerns. The only approved number to call, the IRB ruled, was one at the U.S. university—even though that would require participants to make an international call to someone who didn't speak Japanese! IRB overreach may be one reason that laboratory experiments—particularly impact studies—are becoming less common. Researchers may find it easier to gain IRB approval for online judgment studies than laboratory impact studies. Fortunately, progress is being made on this front; the National Academy of Sciences changed the guidelines for research with human participants, reducing unnecessary IRB oversight over low-risk studies (National Research Council, 2014). Many universities have been slow to adopt these new guidelines, but we hope they do so soon.

Another barrier to conducting experiments—particularly in the laboratory—has inadvertently resulted from concerns about replicability and questionable research practices in psychology. Several reforms have been suggested, and in some cases mandated, to improve scientific practices. All these reforms are well-intended, and many clearly have merit. For example, requiring researchers to describe their methods in detail in supplemental materials, and to make data easily available, have clear advantages, despite the extra time and effort entailed. But we believe the jury is still out on other proposed reforms. As mentioned earlier, pressures to collect large samples come with tradeoffs (Finkel et al., 2017). The question of when to preregister a study—and even whether it is advisable to conduct confirmatory data analyses—is still being debated (e.g., Rubin & Donkin, 2022). Similarly, it is an open question whether "badges" for open science practices, awarded by journals, increase the quality of research (e.g., Rowhani-Farid, Aldcroft, & Barnett, 2020; Schneider et al., 2022). To be clear, we are not advocating that researchers reject these practices. Rather, we point out merely that many open science practices require nontrivial amounts of time and effort by researchers, and that little research is available yet to test whether they work as intended. We can't help but note the irony of requiring such practices prior to being able to provide solid scientific evidence that they are effective, and that they do no harm. Nor can we resist pointing to the analogy of IRB overreach. We are all for procedures that improve scientific practice and encourage researchers to try out such methods for themselves and determine their usefulness for their own work. Just as all researchers should strive to conduct ethical research, independent of the IRB, all of us should strive to conduct robust science and report it honestly. However, we worry that instituting reforms across the board before the evidence is in may be premature and ultimately prove inapplicable (or even inadvisable) in many areas of research. This is unfortunate, particularly if it makes researchers reluctant to conduct the kinds of laboratory experiments we have discussed in this chapter.

We will close on a more optimistic note. In a previous version of this chapter, we expressed concern that basic and applied research in social psychology were functioning in isolation from each other. Happily, that is changing. Indeed, a new subfield of intervention science has emerged that translates principles honed in the social psychological laboratory into effective interventions addressing a wide variety of social and personal problems (Harackiewicz & Priniski, 2018; Walton & Crum, 2022; Walton & Wilson, 2018). It took extensive laboratory research to discover and understand basic processes such as cognitive dissonance, self-affirmation, mindsets about intelligence, the need to belong, and perceptions of utility and value, and then bold attempts to design and test interventions based on these principles—often with great success. Exciting progress is being made on both fronts, and we have great faith that new generations of social psychologists will continue to discover basic principles, using the experimental techniques described in this chapter, and apply these principles to address real world problems.

REFERENCES

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin, 45,* 842-850. https://doi.org/10.1177/0146167218798821

Andrews, F. M., & Robinson, J. P. (1991). Measures of subjective well-being. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 61– 114). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50007-1

Aronson, E. (1966). Avoidance of inter-participant communication. *Psychological Reports, 19,* 238. https://doi.org/10.1111/j.1475-6811.2009.01216.x

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey and E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, pp. 1-79). Reading, MA: Addison-Wesley.

Aronson, E., Fried, C. B., & Stone, J. (1991). Overcoming denial and increasing the intention to use condoms through the induction of hypocrisy. *American Journal of Public Health, 81,* 1636– 1637. https://doi.org/10.2105/AJPH.81.12.1636

Aronson, E., Wilson, T. D., & Akert, R. M. (1994). *Social psychology: The heart and the mind*. New York: HarperCollins.

Aronson, E., Wilson, T. D., & Brewer, M. (1998). Experimentation in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 99-142). New York: Random House.

Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership, and men* (pp. 177-190). Pittsburgh: Carnegie Press. https://doi.org/10.1525/9780520313514-017

Bargh, J.A. & Chartrand, T.L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H.T. Reis & C.M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp.253-285). New York, NY, US: Cambridge University Press.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*, 1-68. https://doi.org/10.1177/1529100619832930

Baumeister, R., Vohs, K., & Funder, D. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on Psychological Science, 2* , 396-403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29,* 1084-1093. https://doi.org/10.1177/0956797617752640

Brehm, J. W., & Cohen, A. R. (1962). *Explorations in Cognitive Dissonance.* New York: Wiley. https://doi.org/10.1037/11622-000

Bruehlman-Senecal, E., & Ayduk, O. (2015). This too shall pass: Temporal distance and the regulation of emotional distress. *Journal of Personality and Social Psychology, 108*(2), 356–375. https://doi.org/10.1037/a0038324

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist, 64*(1), 1–11. https://doi.org/10.1037/a0010932

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365-376. https://doi.org/10.1038/nrn3475

Campbell, D. T. (1957). Factors relevant to validity of experiments in social settings. *Psychology Bulletin, 54,* 297-312. https://doi.org/10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171– 246). Chicago: Rand McNally.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario Symposium* (Vol. 5, pp. 3– 39). Hillsdale, NJ: Erlbaum. https://doi.org/10.4324/9781315802121

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of 65+ million words. *Psychological Science, 32,* 218-240. https://doi.org/10.1177/0956797620963619

Christensen, L. (1988). Deception in psychological research: When is its use justified? *Personality and Social Psychology Bulletin, 14,* 664-675. https://doi.org/10.1177/0146167288144002

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, 8(10), 211037. https://doi.org/10.1098/rsos.211037

Coan, J. A., Schaefer, H. S., & Davidson, R. J. (2006). Lending a hand: Social regulation of the neural response to threat. *Psychological Science, 17(*12), 1032-1039. https://doi.org/10.1111/j.1467-9280.2006.01832.x

Cohen, D. (1977). *Psychologists on psychology.* New York: Taplinger.

Coles, N. A., Gaertner, L., Frohlich, B., Larsen, J. T., & Basnight-Brown, D. M. (2022). Fact or artifact? Demand characteristics and participants' beliefs can moderate, but do not fully account for, the effects of facial feedback on emotional experience. *Journal of Personality and Social Psychology, 124*(2), 287-310. https://doi.org/10.1037/pspa0000316

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342,* 1887-1892. https://doi.org/10.1056/NEJM200006223422507

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings.* Shokie, IL: Rand McNally.

Cooper, J. (2007). *Cognitive dissonance: Fifty years of a classic theory*. Sage Publications. https://doi.org/10.4135/9781446214282

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One, 8,* e57410. https://doi.org/10.1371/journal.pone.0057410

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27,* 100–108. https://doi.org/10.1037/h0034449

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4, Pt.1), 377-383. https://doi.org/10.1037/h0025589

Dawes, R. B., McTavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a common dilemmas situation. *Journal of Personality and Social Psychology, 35,* 1-11. https://doi.org/10.1037/0022-3514.35.1.1

de la Haye, A. (1991). Problems and procedures: A typology of paradigms in interpersonal cognition. *Cahiers de Psychologie Cognitive, 11,* 279-304.

Diener, E. (1994). Assessing subjective well-being: Progress and opportunities. *Social Indicators Research, 31,* 103-157. https://doi.org/10.1007/BF01207052

Diener, E., & Biswas-Diener, R. (2008). *The science of optimal happiness.* Boston: Blackwell Publishing.

Dupuis, M., Meier, E. & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavioral Research*, *51*, 2228–2237. https://doi.org/10.3758/s13428-018-1103-y

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods, 26,* 398–427. https://doi.org/10.1037/met0000349

Ellsworth, P. C. (2010). The rise and fall of the high-impact experiment. In M. H. Gonzales, C. Tavris, & J. Aronson (Eds.), *The scientist and the humanist: A festschrift in honor of Elliot Aronson* (pp. 79-106). New York: Psychology Press. https://doi.org/10.4324/9780203848012

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. In *Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development,* 953-986. https://doi.org/10.1002/9781118489772.ch30

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75– 109). San Diego: Academic Press. https://doi.org/10.1016/S0065-2601(08)60318-4

Ferraro, P. J., & Miranda, J. J. (2017). Panel data designs and estimators as substitutes for randomized controlled trials in the evaluation of public programs. *Journal of the Association of Environmental and Resource Economists, 4,* 281-317. https://doi.org/10.1086/689868

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology, 58*(2), 203. https://doi.org/10.1037/h0041593

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *.Journal of Personality and Social Psychology, 108*, 275–297. https://doi.org/10.1037/pspi0000007

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology, 113,* 244–253. https://doi.org/10.1037/pspi0000075

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517–537. https://doi.org/10.1037/a0023304

Fordyce, M. W. (1988). A review of research on the happiness measures: A sixty second index of happiness and mental health. *Social Indicators Research, 20,* 355-381. https://psycnet.apa.org/doi/10.1007/BF00302333

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., ... & Schneeweiss, S. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation, 143,* 1002-1013. https://doi.org/10.1161/CIRCULATIONAHA.120.051718

Furrer, R. A., Wilson, T. D., & Gilbert, D. T. (in press). An illusion of unfairness in random coin flips. *Journal of Personality and Social Psychology*.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion, 14*, 251.

https://doi.org/10.1037/a0036052

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and laboratory in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*, 847-857. https://doi.org/10.3758/s13423-012-0296-9

Gilbert, D. T. (2006). *Stumbling on happiness.* New York; Knopf.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*, 509-517. https://doi.org/10.1037/0022-3514.60.4.509

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037. https://doi.org/10.1126/science.aad7243

Giner-Sorolla, R. (2025). Changing practices and priorities in social psychological research methods and reporting In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The handbook of social psychology* (6th ed). Situational Press. https://doi.org/10.70400/ZUTF8520

Goldy, S. P., Jones, N. M., & Piff, P. K. (2022). The social effects of an awesome solar eclipse. *Psychological Science, 33*, 1452-1462. https://doi.org/10.1177/09567976221085501

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science, 38,* 193-225. https://doi.org/10.1287/mksc.2018.1135

Göritz, A. S. (2014). Incentive effects. In *Improving Survey Methods* (pp. 339-350). Routledge. https://doi.org/10.4324/9781315756288

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science, 12,* 731-741. https://doi.org/10.1177/1745691617691949

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them.* Penguin Press.

Harackiewicz, J.M. & Priniski, S.J. (2018). Improving student outcomes in higher education: The science of targeted intervention. *Annual Review of Psychology, 69*, 409-435. https://doi.org/10.1146/annurev-psych-122216-011725

Harlow, H. F., & Zimmerman, R. R. (1958). The development of affectional responses in infant monkeys. *Proceedings of the American Philosophic Society, 102,* 501-509. https://www.jstor.org/stable/985597

Hauser, D. J., & Schwarz, N. (2015). It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open.* https://doi.org/10.1177/2158244015584617

Heine, S. J., & Lehman, D. R. (1997). Culture, dissonance, and self-affirmation. *Personality and Social Psychology Bulletin, 23,* 389–400. https://doi.org/10.1177/0146167297234005

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hilbig, B. E. (2016). Reaction time effects in lab-versus Web-based research: Experimental evidence. *Behavior Research Methods, 48,* 1718-1724. https://doi.org/10.3758/s13428-015-0678-9

Holmes, D. S. (1976a). Debriefing after psychological experiments: I. Effectiveness of postdeception dehoaxing. *American Psychologist, 31,* 858-867. https://doi.org/10.1037/0003-066X.31.12.858

Holmes, D. S. (1976b). Debriefing after psychological experiments: II. Effectiveness of postexperimental desensitizing. *American Psychologist, 31,* 868-875. https://doi.org/10.1037/0003-066X.31.12.868

Hopkins, N., & Greenwood, R. M. (2013). Hijab, visibility and the performance of identity. *European Journal of Social Psychology, 43*, 438–447. https://doi.org/10.1002/ejsp.1955

Hsee, C. K., & Ruan, B. (2016). The Pandora effect: The power and peril of curiosity. *Psychological Science, 27*(5), 659–666. https://doi.org/10.1177/0956797616631733

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science, 15,* 809–816. https://doi.org/10.1177/1745691620902467

Kelman, H. (1966). Deception in social research. *Transaction, 3,* 20-24. https://doi.org/10.1007/BF02804546

Killingsworth, M. A., Kahneman, D., & Mellers, B. (2023). Income and well-being: A conflict resolved. *Proceedings of the National Academy of Sciences, 120*(10), e2208661120. https://doi.org/10.1073/pnas.2208661120

Kitayama, S., Tompson, S., & Chua, H. F. (2014). Cultural neuroscience of choice justification. In J. P. Forgas & E. Harmon-Jones (Eds.), *Motivation and its regulation: The control within* (pp. 313–330). New York: Psychology Press.

Kosinski, M. (2025). Using big data. In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The handbook of social psychology* (6th ed). Situational Press. https://doi.org/10.70400/AWLY9440

Latané, B., & Dabbs, J. M. (1975). Sex, group size, and helping in three cities. *Sociometry, 38*, 180–194. https://doi.org/10.2307/2786599

Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology, 10*(3), 215-221. https://doi.org/10.1037/h0026570

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin, 89*, 308–324. https://doi.org/10.1037/0033-2909.89.2.308

Ledgerwood, A. (2018). The preregistration revolution needs to distinguish between predictions and analyses. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606. https://doi.org/10.1073/pnas.1708274114

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the overjustification hypothesis. *Journal of Personality and Social Psychology, 28,* 129 –137. https://doi.org/10.1037/h0035519

Letzter, R. (2021). A teenager on TikTok disrupted thousands of scientific studies with a single video. *The Verge.* Retrieved from: https://www.theverge.com/2021/9/24/22688278/tiktok-science-study-survey-prolific

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science*, 360, 1465-1467. https://doi.org/10.1126/science.aap8731

Linder, D. E., Cooper, J., & Jones, E. E. (1967). Decision freedom as a determinant of the role of incentive magnitude in attitude change. *Journal of Personality and Social Psychology*, *6*, 245. https://doi.org/10.1037/h0021220

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *The Behavioral and Brain Sciences*, *35*, 121. https://doi.org/10.1017/S0140525X11000446

Ma, A., Axt, J., & Kay, A. C. (2019). A control-based account of stereotyping. *Journal of Experimental Social Psychology, 84,* Article 103819. https://doi.org/10.1016/j.jesp.2019.103819

Markey, P. M. (2000). Bystander intervention in computer-mediated communication. *Computers in Human Behavior, 16*(2), 183-188. https://doi.org/10.1016/S0747-5632(99)00056-4

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455. https://doi.org/10.1037/a0028085

Miles, E., & Crisp, R. J. (2014). A meta-analytic test of the imagined contact hypothesis. *Group Processes & Intergroup Relations*, *17*(1), 3–26. https://doi.org/10.1177/1368430213510573

Milgram, S. (1974). *Obedience to authority: An experimental view.* New York: Harper & Row.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science, 62,* 760-775. https://doi.org/10.1111/ajps.12357

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379-388. https://doi.org/10.1037/0003-066X.38.4.379

National Research Council (2014). *Proposed revisions to the common rule for the protection of human participants in the behavioral and social sciences*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18614

Nelson, P. M., Scheiber, F., Laughlin, H. M., & Demir-Lira, Ö. (2021). Comparing face-to-face and online data collection methods in preterm and full-term children: An exploratory study. *Frontiers in Psychology*, 12, 733192. https://doi.org/10.3389/fpsyg.2021.733192

Nisbett, R. E., &Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84,* 231– 2 59. https://doi.org/10.1037/0033-295X.84.3.231

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial feedback effect. *Journal of Personality and Social Psychology, 114,* 657–664. https://doi.org/10.1037/pspa0000121

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600-2606. https://doi.org/10.1073/pnas.1708274114

Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., & Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology*, *6*(1), 17213. https://doi.org/10.1525/collabra.17213

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science,* 349 (6251), aac4716. https://doi.org/10.1126/science.aac4716

Orne, M. (1962). On the social psychology of the psychological experiment. *American Psychologist, 17*, 776-783. https://doi.org/10.1037/h0043424

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology, 120*, 1013. https://doi.org/10.1037/pspp0000358

Paluck, E. L., & Shah, A. K. (2025). Field research: Five approaches to theoretical development in the field. In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The handbook of social psychology* (6th ed). Situational Press. https://doi.org/10.70400/MHNP3692

Paluck, E., Green, S., & Green, D. (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy*, *3*(2), 129-158. https://doi.org/10.1017/bpp.2018.25

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press. https://doi.org/10.1016/S0262-4079(11)62167-2

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-4964-1

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41*, 847-855. https://doi.org/10.1037/0022-3514.41.5.847

Pfattheicher, S, Lazarević, L.B., Westgate, E.C., & Schindler, S. (2021). On the relation of boredom and sadistic aggression. *Journal of Personality & Social Psychology, 121(3)*, 573-600. https://doi.org/10.1037/pspi0000335

Philpot, R., Liebst, L. S., Levine, M., Bernasco, W., & Lindegaard, M. R. (2020). Would I be helped? Cross-national CCTV footage shows that intervention is the norm in public conflicts. *American Psychologist, 75*, 66-79. https://doi.org/10.1037/amp0000469

Plötner, M., Over, H., Carpenter, M., & Tomasello, M. (2015). Young children show the bystander effect in helping situations. *Psychological Science, 26,* 499–506. https://doi.org/10.1177/0956797615569579

Richeson, J. A., & Trawalter, S. (2005). Why do interracial interactions impair executive function? A resource depletion account. *Journal of Personality and Social Psychology, 88*, 934. https://doi.org/10.1037/0022-3514.88.6.934

Roberts, S. (2020, Sept. 2). Sophia Farrar dies at 92; belied indifference to Kitty Genovese attack. New York Times. Retrieved June 27, https://www.nytimes.com/2020/09/ophiaegion/sophia-farrar-dead.html?searchResultPosition=1

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1*, 27-42. https://doi.org/10.1177/2515245917745629

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science, 3,* 176-179. https://doi.org/10.1111/1467-8721.ep10770698

Rosenthal, R., & Lawson, R. (1964). A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats. *Journal of Psychiatric Research, 2,* 61-72. https://doi.org/10.1016/0022-3956(64)90003-2

Ross, L. (in press). Demonstration experiments: Their value, limitations, and relevance to replicability issues. In E. Pronin (Ed.), *Beyond social conflict: Psychological perspectives of Lee Ross*. Oxford University Press.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32,* 880-892. https://doi.org/10.1037/0022-3514.32.5.880

Rossouw, J. E. (2014). Reconciling the divergent findings from clinical trials and observational studies of menopausal hormone therapy for prevention of coronary heart disease. *Seminars in Reproductive Medicine, 32*, 426-432. https://doi.org/10.1055/s-0034-1384625

Rowhani-Farid, A., Aldcroft, A., & Barnett, A. G. (2020). Did awarding badges increase data sharing in BMJ Open? A randomized controlled trial. *Royal Society Open Science, 7: 191818* https://doi.org/10.1098/rsos.191818

Rubin, M., & Donkin, C. (2022): Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests, *Philosophical Psychology*, https://doi.org/10.1080/09515089.2022.2113771

Schachter, S. (1959). *The psychology of affiliation: experimental studies of the sources of gregariousness*. Stanford, CA: Stanford University Press.

Schneider, J., Rosman, T., Kelava, A., & Merk, S. (2022). Do open-science badges increase trust in scientists among undergraduates, scientists, and the public? *Psychological Science,* 33(9), 1588–1604. https://doi.org/10.1177/09567976221097499

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE, 109*, 612-634. https://doi.org/10.1109/JPROC.2021.3058954

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a “direct” replication? Concepts, contexts, and operationalizations. *Social Psychology, 45*(4), 305–306.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology’s view of human nature. *Journal of Personality and Social Psychology, 51,* 515-530.

https://doi.org/10.1037/0022-3514.51.3.515

Sharpe, D., Adair, J. G., & Roese, N. J. (1992). Twenty years of deception research: A decline in participants' trust? *Personality and Social Psychology Bulletin, 18,* 585-590. https://doi.org/10.1177/0146167292185009

Shotland, R. L., & Straw, M. K. (1976). Bystander response to an assault: When a man attacks a woman. *Journal of Personality and Social Psychology, 34,* 990. https://doi.org/10.1037/0022-3514.34.5.990

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., ... & Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. Psychological Bulletin, 144, 343. https://doi.org/10.1037/bul0000128

Sigall, H., Aronson, E., & Van Hoose, T. (1970). The cooperative participant myth or reality? *Journal of Experimental Social Psychology, 6,* 1-10. https://doi.org/10.1016/0022-1031(70)90072-7

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1,* 337-356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. N., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13,* 255-259. https://doi.org/10.1177/1745691617698146

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534. https://doi.org/10.1037/a0033242

Smith, S. S., & Richardson, D. (1983). Amelioration of deception and harm in psychological research: The important role of debriefing. *Journal of Personality and Social Psychology, 44,* 1075– 1082. https://doi.org/10.1037/0022-3514.44.5.1075

Staub, E. (1974). Helping a distressed person: Social, personality, and stimulus determinants. In *Advances in Experimental Social Psychology*, pp. 293-341). Academic Press. https://doi.org/10.1016/S0065-2601(08)60040-4

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54,* 768–777. https://doi.org/10.1037/0022-3514.54.5.768

van Bommel, M., van Prooijen, J. W., Elffers, H., & Van Lange, P. A. M. (2016). The lonely bystander: Ostracism leads to less helping in virtual bystander situations. *Social Influence,11*(3), 141–150. https://doi.org/10.1080/15534510.2016.1171796

Wages, J. E. III, Perry, S. P., Skinner-Dorkenoo, A. L., & Bodenhausen, G. V. (2022). Reckless gambles and responsible ventures: Racialized prototypes of risk-taking. *Journal of Personality and Social Psychology, 122*(2), 202–221. https://doi.org/10.1037/pspa0000287

Walton, G. M., & Crum, A. J. (Eds.) (2022). *Handbook of wise interventions: How social psychology can help people change.* Guilford Press.

Walton, G. M., & Wilson, T. D. (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review, 125,* 617-655. https://doi.org/10.1037/rev0000115

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One, 11,* e0152719. https://doi.org/10.1371/journal.pone.0152719

Westgate, E. C., & Wilson, T. D. (2018). Boring thoughts and bored minds: The MAC model of boredom and cognitive engagement. *Psychological Review, 125,* 689-713. https://doi.org/10.1037/rev0000097

Wicker, A. W. (1969). Attitudes versus actions: The relationship between verbal and overt behavioral responses to attitude objects. *Journal of Social Issues, 25,* 41– 78. https://doi.org/10.1111/j.1540-4560.1969.tb00619.x

Williams, K. D., & Sommer, K. L. (1997). Social ostracism by coworkers: Does rejection lead to loafing or compensation? *Personality and Social Psychology Bulletin, 23*(7), 693-706. https://doi.org/10.1177/0146167297237003

Williams, K. D., Cheung, C. K. T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the Internet. *Journal of Personality and Social Psychology, 79,* 748-762. https://doi.org/10.1037/0022-3514.79.5.748

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science, 1*, 186-197. https://doi.org/10.1177/2515245918767122

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

Wilson, T. D., & Aronson, E., & Carlsmith, K. (2010). The art of laboratory experimentation. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *The handbook of social psychology* (5th ed., pp. 49-79). New York: Wiley. https://doi.org/10.1002/9780470561119.socpsy001002

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 35, pp. 345-411). San Diego, CA: Academic Press. https://doi.org/10.1016/S0065-2601(03)01006-2

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123– 205). Orlando, FL: Academic Press. https://doi.org/10.1016/S0065-2601(08)60311-1

Wilson, T. D., Hodges, S. D., & LaFleur, S. J. (1995). Effects of introspecting about reasons: Inferring attitudes from accessible thoughts. *Journal of Personality and Social Psychology, 69*, 16-28. https://doi.org/10.1037/0022-3514.69.1.16

Wilson, T. D., Lisle, D., Schooler, J., Hodges, S. D., Klaaren, K. J., & LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin, 19* , 331–339. https://doi.org/10.1177/0146167293193010

Wilson, T. D., Reinhard, D., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., Brown, C. L., & Shaked, A. (2014). Just think: The challenges of the disengaged mind. *Science, 345*, 75-77. https://doi.org/10.1126/science.1250830

Wu, J. Q., Horeweg, N., de Bruyn, M., Nout, R. A., Jürgenliemk-Schulz, I. M., Lutgens, L. C., ... & Koelzer, V. H. (2022). Automated causal inference in application to randomized controlled clinical trials. *Nature Machine Intelligence, 4*, 436-444. https://doi.org/10.1038/s42256-022-00470-y

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*. https://doi.org/10.1177/1745691617693393

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111,* 493. https://doi.org/10.1037/pspa0000056

Zhou, S., Page-Gould, E., Aron, A., Moyer, A., & Hewstone, M. (2019). The Extended contact hypothesis: A meta-analysis on 20 years of research. *Personality and Social Psychology Review, 23*(2), 132–160. https://doi.org/10.1177/1088868318762647

ENDNOTES