CHAPTER II

Evaluating Experimental Research

Henry L. Roediger, III and Jeremy K. Yamashiro
Washington University in St Louis

Experimental research is powerful because it generally permits cause and effect statements of relations between variables. The independent variable in an experiment is the one manipulated, whereas the dependent variable is the one measured. If the independent variable has an impact, the level of the variable that is measured depends on the level of the independent variable. Subject variables are differences between people, and they are often of interest in psychological research. However, by definition, they cannot be manipulated because people come to the experiment with the variable assigned to them; thus, causal statements are difficult to make due to potential confounding factors, or the fact that some other variable might be correlated with the one of interest. This can be a problem in all types of experimental research, but it is particularly common when subject variables are examined. The problem of confounding in such research can be minimized by equating subjects on dimensions that are not relevant to the issue under investigation (e.g., age or IQ, if those are not being examined). Finally, control variables are ones that are held constant or randomized across conditions. They may be important, because it can be that whatever effect is observed in an experiment depends on the level of the variables that are held constant. That is, the effect may go away if a different level of the control variable had been used.

Any single experiment must be replicated to be considered valid knowledge. Direct replication, as the name implies, attempts to reproduce an experiment as closely as possible to the original experiment. To advance scientific knowledge, direct replications of new discoveries are critical. Assuming an effect can be directly replicated, the next question concerns boundary conditions of the effect, if any. Systematic replications are experiments in which control variables from the original experiment are permitted to vary a bit to see if the effect can still be found. If the answer is yes, then the experimental effect seems reasonably robust; if not, then the effect is brittle. That is, if even slight changes are made to the original procedure, the effect "breaks" or disappears. Conceptual replications are

those in which large changes are made in the procedure to see if the concept replicates. For example, a new set of materials that seem to have the same properties as the original set might be used to see if some effect generalizes across this major change.

Although replication is critical to science, recent attempts to replicate some well-known studies have failed, which has led to a replication crisis not only in psychology but in all of science. This issue has led to a renewed importance in the eyes of many to emphasizing replications of all sorts, but especially direct replications. Within the last decade, many replication projects have attempted to replicate directly studies that are both famous historically and ones that have been more recently published. The success of these attempts has been less than desired (roughly half replicate robustly), and thus many changes are being urged in scientific practices in general and for psychology in particular.

We have discussed generalizability of research in terms of four factors identified by James Jenkins: Subjects (or participants); materials used in the experiment; dependent measures; and the experimental situation (instructions, the experimental setting, and so on). Effects that generalize across these sets of factors are robust.

We have ended the chapter with a set of 18 critical questions that should be borne in mind while reading and evaluating experimental research. Referring to these questions will help to sharpen critical thinking skills about experimental research.

Introduction

> [T]he application of the experimental method to the problem of mind is the great outstanding event in the study of the mind, an event to which no other is comparable.
>
> Edwin G. Boring (1886–1968)

The author of this quote was one of the great psychologists of the twentieth century, who wrote this line in *A History of Experimental Psychology* (1929, p. 659). Contemporary psychologists take "the psychology experiment" as a given, but it is actually a relatively recent cultural invention. Although fascination with human behavior is doubtless as old as the emergence of *Homo sapiens*, the application of experimental methods to the study of the human mind and behavior is only 160 or so years old. Scientific methods, with heavy reliance on experimental technique, arose in Western civilization during the time of the Renaissance, when great insights and modes of

thought from the ancient Greek, Roman, and Arab civilizations were rediscovered. The seventeenth century witnessed the great discoveries of Kepler, Galileo, and Newton in the physical world. Interest in chemistry and biology arose after the early development of physics. Experimental physiology arose as a discipline in the late 1700s and early 1800s. Still, despite great advances in these fields and despite the fact that scientists of the day usually conducted research in many different fields, no one at that time performed experiments studying living humans or their mental life. The first physiologists and anatomists had mostly contented themselves with the study of corpses. The idea of conducting experiments on mental phenomena in people would doubtless have seemed exotic, if not deemed utterly impossible.

Of course, philosophers and scientists of the time were keenly interested in the mind and mental happenings. The topics of perception, learning, memory, thinking, and reasoning were widely discussed in scholarly writings. Just among British philosophers, John Locke, Thomas Hobbes, George Berkeley, David Hume, and David Hartley all wrote treatises that were concerned with the issues that today preoccupy psychologists. Despite the fact that these men were all aware of the great scientific advances of their time, none of them did experiments to illuminate or test their ideas about the mind. Why? The idea of an experimental science of the human mind had not yet taken hold; no one had yet shown that it could be done. It was not until the period between 1850 and 1900 that bold thinkers turned their experimental techniques to the study of mental life and human behavior.

Consider the following quote from Sir Francis Bacon (1620/2000, p. 143) in making a point about human memory: "If you read a piece of text through 20 times, you will not learn it by heart so easily as if you read it ten times while attempting to recite it from time to time and consulting the text when your memory fails." Give this quote to any competent student of experimental psychology today and it immediately calls forward a hypothesis that could be converted into a psychology experiment. The hypothesis is that learning and memory will be improved during repeated attempts to learn if tests are interspersed with study periods, relative to a condition in which only study periods are given.

Here is a possible experiment: Imagine that passages are created that take about 2 minutes to read. People could be asked to read the passage 20 times (with 2 minutes provided per time) or they could be asked to read the passage, take a test for 2 minutes, read it again, take a test, and so on. The hypothesis predicts that the study-test condition (10 study periods and 10

tests) would lead to better learning and retention than would 20 study periods. If tested a week later, people should show greater retention if they have received 10 study and 10 test trials than if they had received 20 study trials (despite the fact that people tested in the latter condition would have actually studied the material more often). Of course, other possible arrangements are possible too, such as 15 study periods and 5 tests or 5 study periods interspersed with 15 tests. The point here is simply that Francis Bacon made an assertion about memory, probably based on his own experience, that was open to empirical test. However, he did not test his ideas. It took another 300 years for Bacon's idea to be put to experimental test, when Gates (1917) did so. The psychology experiment had not been invented in Bacon's time. Gates showed that Bacon's idea was essentially correct (but see McDermott & Naaz, 2014), and other studies conducted over the years have confirmed the conclusion that testing can be more beneficial to long-term retention than is repeated studying (e.g., Roediger & Karpicke, 2006; Tulving, 1967). Converting hypotheses into experimental tests is at the heart of experimental methods.

The Experimental Method

The heart of the experimental method is straightforward. A hypothesis generated from a theory suggests the relation between two (or more) variables that exist in nature. For example, Bacon's aforementioned hypothesis could be stated this way: Tests of memory interspersed with study periods improve later retention relative to only studying (all other things being equal). A *variable* in an experiment is any factor that can be manipulated or measured. In the experiment outlined here, the number of test periods would be what is called the independent variable, so there might be 0, 5, 10, or 15 tests interspersed among study intervals for this experiment. When there are zero tests, this is a pure study condition, studying the passage twenty times. An *independent variable* is the factor that is manipulated in the experiment; the researcher wants to determine how its manipulation affects some outcome or behavior, the dependent variable. The *dependent variable* in an experiment is what is measured; the name indicates that, in most circumstances, variation in the measure of interest will depend upon the level of the independent variable. Of course, this is not always the case, because manipulation of the independent variable may not affect the dependent variable. The hypothesis under test may be wrong or, alternatively, the independent variable may not be manipulated over a wide enough range to affect the dependent variable. (In

the sample experiment outlined here, the dependent variable would be how well the passages are recalled on a delayed test a week later.) How might "recall of prose" be measured? The usual method is for passages to be divided into idea units (small units of text that constitute an idea, as judged by people rating the text). Therefore, the dependent measure would be the number of idea units recalled or the percentage of idea units recalled.

Another set of factors in an experiment is called *control variables* (although the name is a bit of a misnomer). In our version of Bacon's hypothesis, the phrase "all other things being equal" appeared in parentheses and these "things" are the control variables. Control variables are factors that the experimenter could manipulate, but instead holds constant as much as possible. If they cannot be held constant, they are randomized across conditions. The idea behind an experiment is to determine what effect the manipulation of the independent variable has on the dependent variable. It is critical to hold all other conditions as constant as possible to ensure that if an effect is found on the dependent variable, it was caused by the independent variable. If other variables are allowed to change over conditions, then they might be causing change in the dependent variable and not the independent variable of interest. If some other factor varies along with the independent variable, the experiment is said to be *confounded*, because any effect observed on the dependent variable may have been caused by the independent variable or the other, confounding variable. A *confounding* exists whenever some other factor varies with, or is correlated with, the independent variable of interest. The problem of confounding undermines the rationale for experimental research, so great effort and care are taken in experimental research to hold other factors constant so as not to permit confoundings. However, this is sometimes difficult to accomplish because when a researcher manipulates what seems to be one variable, that variable may actually be composed of several features (unbeknownst to the researcher). Therefore, the researcher might believe that Feature *A* is causing the experimental effect, but Features *B* and *C* vary with *A*. Further research might show that Feature *C* is actually causing the effect and *A* is not.

Control variables may not seem so important to research, because they are the features of the experiment held constant. However, they are actually critical, because the level at which the control variables are held constant may determine the outcome of the research. Suppose, for example, that an experiment is conducted testing some hypothesis about human memory by having various groups of college students learn lists of words. Some independent variable is manipulated and the number of words recalled

(the dependent variable) is measured. Control variables in this experiment are the types of subjects (college students) and materials (the word lists). These factors may or may not turn out to be important. If later research is done with elementary school children and the results turn out differently, then clearly the control variable (types of subjects used in the original research) was important. Similarly, if different effects occur with materials besides words lists, such as prose passages, then the type of materials also would turn out to be an important variable. We return to this issue later in the chapter when we consider generality of experimental research. The point here is that selection of control variables (or features of the experiment that are not varied) may have as great a consequence in the long run as the features that are varied.

Between-Subjects and Within-Subjects Designs

Another critical decision in designing experiments is whether to use different sets of people (or other animals) in the various experimental conditions or the same people (or animals). This constitutes a difference in using between-subjects or within-subjects experimental designs. In a *between-subjects design*, a different group of subjects is assigned to each level of the independent variable. So, in the example used here, one group of subjects would be assigned to study the passage twenty times and a different group of subjects would study the passage ten times and be tested ten times in an alternating sequence. Then both groups would be tested a week later. But wait – hasn't this experimental design produced a problem, a factor that differs between the two conditions and so is a potential confounding factor? Yes, that is so – different groups of people are being tested in the two conditions, so how do we know that any difference we find might not be a difference in level of ability between the two groups of subjects? A critical factor in between-subjects experimental designs is that people (or animals) must be *randomly assigned* to conditions (or some other measures must be taken to ensure that they are equal, on average, in ability and other characteristics). For example, when a new subject appears at the laboratory to be tested, a coin flip (or a random number table or some other means) should be used to assign the person to either the pure study or to the study-test condition. This step should ensure that the two groups of people are, on average, about the same in ability and in other qualities. Thus "people in the two conditions" would not literally be held constant as a control variable, but because any differences between subjects would presumably be small ones caused by random assignment, any variation

observed in the dependent variable could safely be attributed to the independent variable and not to differing levels of subjects' abilities. Of course, it helps to have rather large numbers of randomly assigned people to build power in the experiment; any quirk that one subject might have (an especially good or poor memory) would then be minimized.

Another type of between-subjects design is the *matched groups design*, in which some relevant ability of people is measured before the experiment. Then subjects are assigned to groups in the experiment so that they are matched on the relevant dimension. For example, if middle-aged and older adults were compared in an experiment on memory or some other cognitive ability, they might be matched on years of education or on level of vocabulary (a proxy for verbal IQ). In this way, even though different groups of people are tested, the researcher can be relatively assured that there are no important intellectual differences between the groups.

Now let us consider the *within-subjects experimental design* in which the same individuals serve in all experimental conditions. For example, in our experiment on testing, on the first day of the experiment the subjects would study one particular passage twenty times and then, after doing that, they would study a second passage ten times with ten tests interspersed. A week later they would come back and receive a test on both passages. Thus, the two conditions would be compared with the particular subjects participating in the conditions held constant. Although the participants do not differ, other complications are introduced in the within-subjects design. For one thing, there are now two passages and not one (as in the between-subjects design). Does that matter? It probably does, but there are several ways of making sure the passage type is controlled. One strategy is to pretest both passages and make sure they can be read and recalled at about the same level on an immediate test (i.e., the passages are matched). A second strategy is to use both passages (call them *A* and *B*) equally often in both conditions, so that passage *A* is included as often in the pure study condition as it is in the study-test condition across the subjects in the experiment. Using these strategies can convert a problematic situation (two different passages in the two experimental conditions) into one that is well controlled. The trick is making the type of passage a control variable.

The within-subjects design produces other complications, too. One is *practice effects*, or the fact that the subjects will be participating in both conditions and thus practice on one condition might affect how they perform in the next condition. Thus, practice can introduce a confound with the independent variable of interest. The way to minimize this problem is *counterbalancing* the order of conditions across subjects. That

is, if the conditions are *X* and *Y*, half the subjects will get them in the order *X* then *Y*, whereas the other half would get them in the order of *Y* then *X*. Thus, on balance, each condition would be tested at the same (average) stage of practice, and therefore stage of practice would not be confounded with the variable of interest. Of course, counterbalancing is easy when there are two conditions in an experiment, but in certain types of research there may be many conditions and counterbalancing becomes much more difficult. Various strategies for counterbalancing can be found in textbooks on experimental design (e.g., Kantowitz, Roediger, & Elmes, 2015).

A problem in within-subject designs that is more difficult to overcome is the *differential carryover effect.* Unlike general practice effects, in this case participation in one condition can greatly change performance in the other condition. Suppose, for example, that a researcher is interested in whether creating mental images is a good strategy for memorizing, relative to simple repetition. If a within-subjects design is used, the order of the two conditions must be counterbalanced, with subjects instructed to learn materials in one condition by rehearsing (mentally repeating) them and in the other by forming mental images that would depict the materials (e.g., if they had to learn the pair of words "clock–tree," they might imagine a giant clock hanging from a tree). However, if subjects are first tested in the imagery condition and they discover that it works really well (which it does), then when they are switched to the repetition condition, they might still use imagery, in the interest of performing their best. (The problem would probably be less severe for subjects tested in the repetition-then-imagery order, because they would be less likely to carry over the repetition strategy, as it is less effective.) In this case, counterbalancing of conditions will not eliminate practice effects, because there might be differential carryover from the imagery to the repetition condition that would create a confounding in the experiment. Of course, in this kind of experiment, the researcher can always just examine the half of the subjects given each treatment first. That is, because half the subjects would get the repetition-then-imagery condition and the other half the imagery-then-repetition one, the experimenter could examine the first condition for both groups. In doing this, the researcher would essentially be treating the study as having a between-subjects design with one group of subjects in one condition and the other group in the other condition. If the results do not differ between the first condition and the overall experiment, the experimenter can conclude that there were no serious carryover effects.

Which type of experimental design is generally the best, between-subjects or within-subjects? There can be no general answer to this question,

because the design selected to answer a particular question can depend on many factors. An advantage of within-subjects designs is that the same people (or animals) are used in both conditions, which often reduces variation in performance that is due to having different people in the groups. That is, even if people are randomly chosen to be in the two conditions in a between-subjects design, the groups may differ in small ways and this feature is eliminated from within-subjects designs. Put another way, the power of an experiment (the ability to detect an effect of the independent variable if there really is one) is usually greater in within-subjects designs, too, because the same people are tested in each condition and thus variability among people is minimized. However, if serving in one condition can affect performance in the other condition, then a between-subjects design may be preferred. In the hypothetical experiment comparing an imagery to a repetition strategy for improving memory, a between-subjects design would probably be best. In general, if one condition being tested is likely to influence the other conditions in the experiment greatly, then a between-subjects design is preferred.

Subject Variables

We have distinguished among independent variables, dependent variables, and control variables. We need to introduce one more type of variable, the *subject variable*, because much psychological research uses this sort of variable. A factor is referred to as a subject variable when individual differences among people are examined on some task or set of tasks. It is somewhat like an independent variable – its effect on the dependent variable is the factor of interest in the experiment – but there are also important differences between independent and subject variables. For example, a researcher may be interested in some behavior of younger children (3–5 years of age) or older children (7–9), or the investigator may be interested in people with high IQs (scores of 120 and up) or those in the normal range (85–115), or the interest may be in older adults (65–90 years of age) and younger adults (20–30), or people who have an anxiety disorder (say, phobias to snakes and spiders) and those who do not have such phobias, and so on. The study of differences among people is a staple of psychology. However, this is a variable that cannot be manipulated like a true independent variable and, by definition, subject variables cannot be randomly assigned to conditions – people are assigned by nature to the variable. The great danger with subject variables is that some other factor might be correlated with the factor of interest and therefore

introduces a confounding in the experiment. Because of this problem, great care is taken in such research to try to match people in the various conditions, as already described. For example, in a study of age difference between young, middle-aged, and older adults, subjects in the three groups would be matched as closely as possible on at least several features, which would typically include education, eyesight (corrected to normal), self-reported health, and often vocabulary (as assessed on standardized tests). Matching in this way reduces the risk that the findings from the study may be due to some factor other than the variable of interest, in this case age.

These considerations provide a summary of critical features of experimental methods. To recap briefly, independent variables are those that are manipulated; dependent variables are those that are measured; control variables are features that are held constant; and subject variables are features of people or animals assigned by nature, so that when they are examined care must be taken to match the individual on other characteristics as much as possible. Experimental methods seek to study effects of one or several variables on some behavior while holding others constant.

A Sample Experiment

We present a sample experiment that we use to discuss the critical evaluation of experimental research. The experiment is on the issue of false memories. A false memory occurs when a person remembers an event differently from the way it actually happened or, in the most extreme case, remembers an event that never happened at all. Usually we do not know when our own recollections are false because we believe them; if someone else has a different recollection, we tend to believe our own and assume the other person is mistaken. However, consider the following anecdote from Jean Piaget (1962. pp. 187–188), the great Swiss psychologist, about a cherished memory from his childhood:

> There is also the question of memories which depend on other people. For instance, one of my first memories would date, if it were true, from my second year. I can still see, most clearly, the following scene, in which I believed until I was about fifteen. I was sitting in my pram, which my nurse was pushing in the Champs-Élysées, when a man tried to kidnap me. I was held in by the strap fastened round me while my nurse bravely tried to stand between me and the thief. She received various scratches, and I can still see vaguely those on her face. Then a crowd gathered, a policeman with a short cloak and a white baton came up, and the man took to his heels. I can still see the whole scene and can even place it near the tube station. When I was

> about fifteen, my parents received a letter from my former nurse saying that she had been converted to the Salvation Army. She wanted to confess her past faults, and in particular to return the watch she had been given as a reward on this occasion. She had made up the whole story, faking the scratches. I, therefore, must have heard, as a child, the account of this story, which my parents believed, and projected into the past in the form of a visual memory.

Psychologists interested in this issue have developed laboratory paradigms to create and study various types of false memories. Many different paradigms have been developed (see Roediger & Gallo, 2002, for a review). Here we consider one straightforward paradigm developed by Roediger and McDermott (1995), which was based on earlier work by Deese (1959) and is now known as the converging associates or DRM paradigm (for Deese-Roediger-McDermott). The basic paradigm involves presenting lists of related words such as *door, glass, pane, shade, ledge, sill, house, open, curtain, frame, view, breeze, sash, screen,* and *shutter*. After hearing the list presented once (at a rate of about 1.5 seconds per word), subjects recalled the list on a blank sheet of paper by writing down the presented words. They were warned against guessing and told to be as accurate as possible. Usually students recalling lists of words are highly accurate and make few errors, especially on immediate tests. However, that was not the case in this experiment.

The basic finding from the Roediger-McDermott experiment was that the subjects were highly likely to recall or to recognize a particular associated word that was not in the list (*window* in the case of the list given here). The fifteen presented words were taken from word association norms; when subjects were given the word *window* and asked to write down the first word that came to mind, the fifteen words listed here were the most probable responses. When subjects recall *window* just after the list is presented, they are (according to the definition given here) having a false memory: They are recalling an event (the occurrence of a word in a list) that did not happen. Of course, this type of laboratory false memory probably arises for different reasons from those for Piaget's false memory recounted in the previous paragraph, but in both cases, there is a firm recollection of an event that never occurred. The basic idea as to why DRM false memories occur is that the presented words are associated with other words (like *window*) and these associations are activated when people hear the list (Underwood, 1965). Such implicit associative responses (which the subject may or may not become conscious of during the study presentation) give rise to the DRM false memory effect, because during the test the subjects have difficulty in distinguishing activation that arose from actually

encountering presented words (*frame, screen*) from that which arose from activation spreading through the cognitive system (*window*). They often judge strongly activated words such as *window* as having actually occurred and report them in a recall test or endorse them on a recognition test (see Gallo, 2010, and Roediger & Gallo, 2016 for reviews).

Let us consider some specifics of one of Roediger and McDermott's (1995) experiments to help us critically evaluate it. They developed 24 lists like the one already given here, all containing 15 words that were associates of a particular word (the critical item) that was not presented. In the experiment, students heard 16 lists one word at a time for 1.5 seconds per word. After 8 of the lists they recalled as many words as they could for 2 minutes. The rationale here was to determine the effect of whether or not a list was recalled on a recognition test to be given later, but of course false recall just after the list could be examined, too, on the 8 recalled lists.

After studying the 16 lists and recalling half of them, students were given a final recognition test at the end of the experiment. In this test, 96 words were presented and students had to decide whether or not each one had been previously presented (was it old or new?). Further, if they judged that the word was old (that is, had been presented), they were asked if they could remember specific details about the moment the word occurred in the list (a Remember judgment) or if they just knew it had been presented but could not actually remember any details about its occurrence in the list (a Know judgment). This Remember-Know procedure was developed by Tulving (1985) to study states of awareness accompanying retrieval of past events. Of the 96 words on the recognition test, half had been presented and half had not. The 48 presented words included 3 each from the 16 studied lists. The 48 non-presented words (called *lures* or *distractors* in the context of recognition testing) consisted of 16 critical items from the presented lists (the words like "window" that had been used to generate the lists) and 32 new words were taken from the 8 lists that had not been studied. Recall that 24 lists were developed but only 16 were presented, so the other lure items on the recognition test were the 8 critical items and the 24 list items from the lists that *had not* been presented. The reason for having these items is to examine the general false alarm (or false memory) levels on the test when the relevant lists had not been studied.

On the initial recall tests given for 8 lists just after they had been studied, students recalled 62 percent of the list words. However, they also recalled the critical (nonstudied) items from those lists 55 percent of the time! That is, the words like "window" that were not presented were recalled with nearly the same probability as the words like "glass" and the others that were

presented. Keep in mind that this occurred despite the fact that the tests were given immediately after presentation and that the subjects were instructed to be sure to write down only words that they had just heard. We will not discuss the results of the recognition test in detail here, but in general the results showed very high levels of false recognition of words like "window" that were implied by the list but not presented. In addition, participants reported *remembering* – rather than *knowing* – having seen the critical lures at similar rates as they reported for actually presented items. This outcome thus truly represented false remembering. The Roediger and McDermott (1995) results show that even in the simple task of recalling a list of words, people can still suffer from false recollections. These results have been studied and debated for the past twenty-five years, so we use this paper as a target to critically evaluate experimental research.

Generality and Limitations of Experimental Research

Critics of experiments like to point to their artificial nature and their limitations. The issue of false memories arises in many critical situations: in congressional testimony, in legal settings, in recountings of important meetings of interest to historians or other scholars, or eyewitness testimony of all sorts. How accurate are recollections? Can what we learn in laboratory paradigms inform larger issues outside the lab?

Discussing these issues gets to the heart of reasons for experiments, both their benefits and their limitations. The benefit of an experiment is to isolate one factor or several factors and hold others constant to examine the effect on the critical measures of interest. The drawback is that in creating a simple setting to isolate one factor, we may reduce the ability to generalize the result back to a complicated setting in which many other factors vary willy-nilly. Of course, this issue is not unique to psychological research but occurs in all types of research. If massive doses of some substance (say, saccharin) are shown in controlled experiments to cause cancer in laboratory mice, should the substance be banned from human consumption? Can we generalize across a different species and a different dose?

The DRM false memory paradigm can be (and has been) criticized as artificial and of little relevance to the development of false memories outside a laboratory setting (Freyd & Gleaves, 1996). However, often laboratory conditions (being carefully constructed and holding other events constant) can actually make it more difficult to observe a particular result. Experiments are conducted to test hypotheses and to determine causality, which are different goals from immediate generalization

(Mook, 1983). In discussing this issue, Roediger and McDermott (1995) made this comment:

> A critic might contend that because these experiments occurred in a laboratory setting, using word lists, with college students, they hold questionable relevance to issues surrounding more spectacular occurrences of false memories outside the laboratory. However, we believe that these are all reasons to be more impressed with the relevance of our results to these issues. After all, we tested people under conditions of intentional learning, with very short retention intervals, in a standard laboratory procedure that usually produces few errors, and we used college students – professional memorizers – as subjects. In short, despite conditions much more conducive to veridical remembering than those that typically exist outside the lab, we found dramatic evidence of false memories. When less of a premium is placed on accurate remembering, and when people know that their accuracy in recollecting cannot be verified, they may even be more easily led to remember events that never happened than they are in the lab. (Roediger and McDermott, 1995, p. 812)

The issue of artificiality of experimental research is a difficult one, and certainly researchers should strive as hard as they can to capture the important aspects of a phenomenon of interest and to bring them into the lab. Experiments are designed to provide internal validity of a result; internal validity is whether the independent variable affected the dependent variable. Is the cause and effect conclusion being drawn from the experiment valid (true)? Experiments are usually high in internal validity. External validity refers to generalizability to other settings. Does the effect observed in an experiment generalize to other settings?

The issues of internal and external validity are critical in research, but there is a priority to these considerations that may not be obvious. Internal validity of research must be established before one can begin to worry about external validity (Mook, 1983). Stated another way, one must have a secure finding (manipulating *X* causes *Y* to vary) before it is even worth worrying about whether this finding occurs in other settings outside the lab. Banaji and Crowder (1989) argued that experimental research (with tight controls) is the best way to guarantee that research is potentially generalizable. Thus, rather than conducting research in "natural" settings in which many factors vary uncontrollably, researchers usually must develop careful laboratory methods to establish firm findings before asking whether these findings can be generalized. Scientists have no inherent fascination with artificial settings, but rather they create these settings as a means to the end of providing conclusions with internal validity.

Assume that scientists have conducted an experiment and obtained a result that they and other scientists deem to be interesting and important. What are the next steps? The critical first step is always replication of the result: Can other researchers conduct the same experiment and get the same result? Of course, the original scientists may already have replicated their finding, but the critical test is whether others, using the same procedures, will find it as well. A *direct replication* refers to performing the experiment in as similar a way as possible in attempting to repeat the result. Although on the face of it we should expect direct replications to be fairly common and easy, in fact recent research has indicated that many fewer studies replicate than would be expected.[1] More on this issue follows below. The next step is *systematic replication*; this term refers to manipulation of all sorts of small variables that, if the phenomenon under consideration is robust, should not matter (Roediger, 2012). For example, in a memory experiment, what college students are from, whether the pictures in the experiment are presented at a 4-second or a 7-second rate, whether people are tested at 2 p.m. or 4 p.m., should generally not matter. The issue is whether a finding is robust across many slight variations. If the phenomenon fails this test – if it can be directly replicated but if even small variations in the procedure makes the effect disappear – the finding is brittle. It is also probably not very important if even small changes in the procedure make it disappear. However, if the experimental effect survives these variations, it is considered robust, at least across small changes in the procedure. Finally, the third type of replication, conceptual replication, is the last important step. A *conceptual replication* refers to seeking the same general pattern of results but using different methods from the original procedure. Can the basic concept of the experiment be replicated? Will the experimental effect survive when the independent variable is manipulated in a different manner or when the measure (the dependent variable) is somewhat different? If a conceptual replication experiment obtains the same results as the original experiment, then the phenomenon has at least some generality. If the effect is not found, then the investigator may have found *boundary conditions* for the phenomenon, or variables beyond which the experimental effect will not generalize. Establishing boundary conditions is quite important in many contexts, both for developing

[1] In a previous edition of this chapter (Roediger & McCabe, 2007, p. 27), it was stated that "We hazard the guess that most experimental results in psychology can be directly replicated (although there are some notable exceptions to this claim)." This guess has since been challenged.

theories of the phenomenon and for practical purposes. For example, if huge doses of saccharin cause cancer in mice but small doses do not cause cancer in human beings, then the generalization that "saccharin causes cancer" does not hold over important conditions.

In recent years, the question of replications, direct, systematic, or conceptual, has become a central concern in the community of scientific researchers, so much so that many have pronounced a "crisis of confidence" (Pashler & Wagenmakers, 2012). When we evaluate an experimental finding, particularly a surprising one, a critical question should be, if someone else ran this experiment again, would we see the same results? The philosopher of science Karl Popper placed such replicability at the center of the scientific endeavor:

> Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested – in principle – by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence,' but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable (Popper, 2002, p. 23).

If an effect does not replicate, we may be justified in questioning whether or not it is real, even if it was deemed "statistically significant" in the original publication. It is therefore particularly troubling if "effects" that are reported in one paper cannot be reproduced by other competent research teams that are closely following the original procedure. In this case, we would say that the effect cannot be directly replicated. Several failures to replicate famous studies have brought attention to the need to critically evaluate the likelihood that any observed finding is true.

To explain the events that brought about the replicability crisis, we take a brief historical detour. For some time, methodologists had warned that common practices in the collection and analysis of data were likely to produce effects that were actually false (e.g. Ioannidis, 2005). However, it was an article published in the *Journal of Personality and Social Psychology* that helped to kick off what has become known as the replicability crisis. In 2011 Daryl Bem, a prominent social psychologist at Cornell University, published a paper in which he claimed to show "anomalous retroactive influences on cognition and affect" (2011, p. 407). In more straightforward language, he claimed to show that humans have extrasensory perception; they can feel and be influenced by events that are about to happen. In two of

his experiments, for example, Bem inverted the usual testing effect paradigm. In the ordinary experiment, participants learn a set of words, retrieve some subset of them and re-study others, then do a final memory task, often a free-recall (e.g. Roediger & Karpicke, 2006). Retrieved items are usually recalled better than items that have not been retrieved, and better than items that have simply been re-studied. Bem's participants, however, underwent the free-recall of all learned words *prior* to the selective retrieval practice phase. In a shocking – some might say impossible – finding, Bem showed that words that were *about to be* retrieved on the second test were recalled better on the initial recall task, despite the fact that any additional retrieval had not yet actually occurred. He argued that this was evidence for precognition or *psi*, an ability to sense the future. In this case, he claimed his subjects were able to sense what condition (test or study) the words would receive in the future. The same sort of "retroactive influence" was seemingly shown in a variety of other ways, across nine experiments.

Bem's paper on extrasensory perception was, with little risk of hyperbole, a lightning rod, because no one believed the outcome despite his seemingly careful methods. If Bem could produce statistically significant results seeming to indicate abilities that were generally agreed to be impossible and which were therefore likely unreal, could other published results be illusory? What other apparently well-supported effects might also be false? With renewed vigilance, large teams of researchers began coordinating attempts to replicate many well-known and widely accepted effects that seemed to be based on one or two studies. For many of these experiments, and especially many in social psychology, these collaboratives soon reported that they could not produce even a direct replication. Large-scale failures to replicate brought many accepted phenomena into question, for example: *social priming* – that unscrambling sentences related to the elderly made college-age participants walk more slowly (e.g. Doyen et al., 2012; Klein et al., 2014); *ego depletion* – the idea that willpower is a limited resource, and exerting willpower in one task will decrease the amount of willpower available for another task (Hagger et al., 2016); *stereotype threat* – that in groups with negative academic stereotypes, fear of conforming to these stereotypes produces anxiety, which is the true cause of group differences in performance (Ganley et al., 2013); and *power posing* – the apparent finding that adopting physical "power poses" increases hormone levels related to dominant behaviors (Ranehill et al., 2015). In a letter to researchers working in the beleaguered field of behavioral priming, the Nobel Prize winning psychologist Daniel Kahneman (2012) warned of a "train wreck looming," and he exhorted

these researchers to take proactive measures to address the impending crisis of confidence in their field.

Factors Likely to Produce False Effects

Why were so many studies failing to replicate, in the sense of direct or systematic replication? One explanation may have been that the effects reported were a result of Type I errors, or the statistical appearance of an effect when in fact no such effect existed. A small but vocal minority of researchers had earlier expressed concerns that common practices in psychological research increased the likelihood of making such errors. For example, Ioannidis (2005) proposed five rules of thumb for judging when a statistically significant research finding is probably untrue:

1) The sample size (or the number of observations) is small;
2) The size of the effect is small;
3) Tests of the data are numerous and exploratory, versus targeted and confirmatory;
4) The number of researcher degrees of freedom (discussed below) are high;
5) External incentives to get significant results are high (as when funding of future research depends on publication, and publication depends on statistically significant findings); and
6) A field is "hot," that is, when a larger number of research teams focus on an issue and compete to produce impressive results.

Shortly after Bem published his controversial paper on precognition, Simmons, Nelson, and Simonsohn (2011) published an article demonstrating how easy it was to get effects that were statistically significant, yet entirely spurious. To do so, they introduced the concept of *researcher degrees of freedom*. Researcher degrees of freedom refer to the choices scientists must make when collecting and analyzing data: how many participants to collect, when to stop data collection, if and when to exclude data, and which variables to include in an analysis, among others. When researcher degrees of freedom are high, the researcher has more flexibility to analyze the data in a variety of different ways before stumbling upon results that happen to attain statistical significance. Frequently, such exploratory analyses result in a practice now known as HARKing – Hypothesizing After Results are Known. In HARKing, a statistically significant finding is used to justify a hypothesis that has been created to explain the result. That is not necessarily bad when the researcher explicitly

indicates that the results were unexpected and uses the surprising finding as a spur for future research; however, some researchers were presenting the hypothesis (derived after the fact) in the introduction of their papers as if it had been their hypothesis all along, prior to data collection, and that the results had confirmed their idea (Kerr, 1998; Bones, 2012). Simmons, Nelson, and Simonsohn argued that this practice of HARKing undermined the legitimate scientific process.

In another way of exploiting researcher degrees of freedom, researchers may continue testing participants until a test reaches statistical significance, and then stop (in case testing more people would make the effect disappear). Although it was once widely believed that large effects in small samples provided convincing evidence that an effect was robust and strong, it is actually the case that spurious effects are more likely to appear in small samples. Such attempts to manipulate the statistical significance of a test in these ways have been termed *p*-hacking, motivated as they are to produce a test with a *p*-value below the critical threshold of .05 (Head et al., 2015). *P*-hacking occurs when the attempt to create statistical significance uses the dubious practices just discussed, along with others that are outside the reach of this chapter.

Scientifically dubious massaging of the data is not the only route to false effects. Indeed, having many researcher degrees of freedom in a study virtually guarantee that false effects will emerge. Each potential set of choices during study design, data collection, and analysis, constitutes a choice pathway in what Gelman and Loken (2013) called a "garden of forking paths." Each path may be innocent and justifiable, but in the aggregate, multiple potential paths increase the likelihood of discovering a statistically significant effect simply by chance. This inflation of the Type I error rate – the probability of "finding" an effect that is not actually there – may occur even if researchers are not intentionally "fishing," HARKing, or *p*-hacking. The mere existence of many researcher degrees of freedom in a study almost guarantees statistically significant but false findings. "Forking paths" and many researcher degrees of freedom represent a royal road to Type I errors, or false findings. Such issues occur much more frequently in large correlational studies with many variables than in true experiments, but the problems exist in both correlational and experimental research.

Reproducibility Projects

In response to this crisis of confidence, several large-scale replication projects were launched as researchers attempted to gain a sense of how many false positives there might actually be in the published literature. In 2011 Brian

Nosek from the University of Virginia and collaborators initiated the Reproducibility Project: Psychology. A team of 270 researchers selected 100 studies to replicate, all drawn from three of the top psychology journals: *Psychological Science, Journal of Personality and Social Psychology*, and the *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Published in 2015, the outcome of this project was disheartening (Aarts et al., 2015). Only 36 percent of the replication studies showed statistically significant effects at $p < .05$. There were sub-disciplinary differences. Using the *p*-value criterion, 51 percent of cognitive-psychology studies replicated; 26 percent of social psychology studies did so. In both subfields, however, effect sizes in the replication studies were, on average, half the size of those originally reported. The Social Sciences Replication Project was another such large-scale attempt at replication. As part of this project, Camerer et al. (2018) selected twenty-one social science studies published in *Nature* and *Science* between 2010 and 2015. Using sample sizes that were on average five times larger than the original studies, they found that effect sizes in the replications were half the size of those reported in the original studies, similar to Aarts et al. (2015). Camerer et al. (2018), however, were somewhat more successful at replicating their target studies; 62 percent of the replications had a statistically significant effect in the same direction as the original study. Interestingly, when Camerer et al. (2018) showed the twenty-one studies to a sample of peers who predicted whether the studies would replicate, people could predict with some accuracy which studies would replicate and which wouldn't.

Needless to say, the research community was unsettled to see replication rates between 36 percent and 62 percent for papers published in the best journals, and public confidence in psychological science faltered. (However, it is clear that the same problems exist in other scientific fields such as cancer biology and probably any field that relies on statistical inference.) In response, psychologists have begun reforming research standards; some of these reforms have become widely adopted, others have yet to do so (Spellman, 2015). Some of the suggested reforms include:

1.) Replacing or supplementing null hypothesis significance testing with either (a) the "new statistics," which emphasizes precise estimates of differences rather than binary significant/non-significant outcomes (Cummings, 2012) or (b) Bayesian analyses, statistical techniques that also allow for probabilistic estimates of how likely an effect is. Such analyses are also informed by prior beliefs in the likelihood of the effect; more convincing evidence is needed for effects that seem highly unlikely ahead of time.

2.) Requiring formal power analyses to determine adequate sample sizes. Sufficient statistical power reduces the risk that an observed effect is spurious. Determining sample size ahead of time also removes the temptation to *p*-hack by stopping data analysis if a significant *p*-value is reached.
3.) Pre-registering hypotheses and planned analyses. This reduces the tendency to hypothesize after the results are known or HARK. Some journals are offering publication of studies that were pre-registered regardless of outcome. This step partially resolves the "file drawer" problem where large numbers of failures to replicate are never published and so are invisible to readers of the scientific literature. Pre-registration also serves to eliminate or tightly control researcher degrees of freedom, because the methods of analysis are specified ahead of time. Of course, exploratory analyses can still be performed, but then they can be identified as exploratory.
4.) Data Sharing. The open sharing of data increases transparency and allows for other researchers to create large analyses (called meta-analyses) across many data sets. The Open Science Framework has become a prominent platform for accessible data banking. Much additional information relevant to replication can be included (e.g., all the materials, the scripts for analyses).
5.) Full descriptions of methods. Some journals have removed word count limits in methods sections; being able to specify methods precisely helps other researchers make good faith attempts at replication.

These revised practices help us to have greater confidence in a particular experimental outcome, or at least help us in clarifying the extent to which we should be uncertain about it. As discussed above, before researchers can attempt systematic replications or explore an effect's boundary conditions, they have to trust that an effect is directly replicable. Confidence in the reality of an effect requires direct replications by independent research teams. Once the basic stability of an effect in a specific context has been adequately established through both direct and conceptual replications, researchers can expand to questions of generalizability. Of course, certain findings about social conditions (e.g., will people intervene in a crisis?) might not replicate because the phenomenon genuinely does change over time. This points for the need to establish boundary conditions on phenomena by performing direct and systematic replications over time.

Jenkins's Approach to the Generalizability of Research

James Jenkins (1979) provided a useful way to think about issues of generality in research. Although Jenkins was concerned with memory research, the points he made apply to all types of research and can therefore be generalized to all areas of psychology. Jenkins pointed out that any single experiment or finding – even if it can be directly replicated with ease – should be considered in the context of the factors that were held constant but that could have potentially been manipulated. That is, every experiment occurs in the context of control variables or factors that were held constant in the research, but that could have been varied. To what extent is the phenomenon of interest determined by the setting of the control variables? Posing the question in a different way, when other variables are manipulated, will the experimental effect still occur under these new conditions? This is one way of thinking about the issue of conceptual replications and generalizability of research.

Jenkins's basic idea of *contextualism* is represented in Figure 11.1, in what he called a tetrahedral model, where it can be seen that four sets of factors are considered. (The model gets its name from the fact that if all the vertices are connected, the resulting shape is a tetrahedron.) One dimension is the type of *subjects* tested: children, white rats, college students,

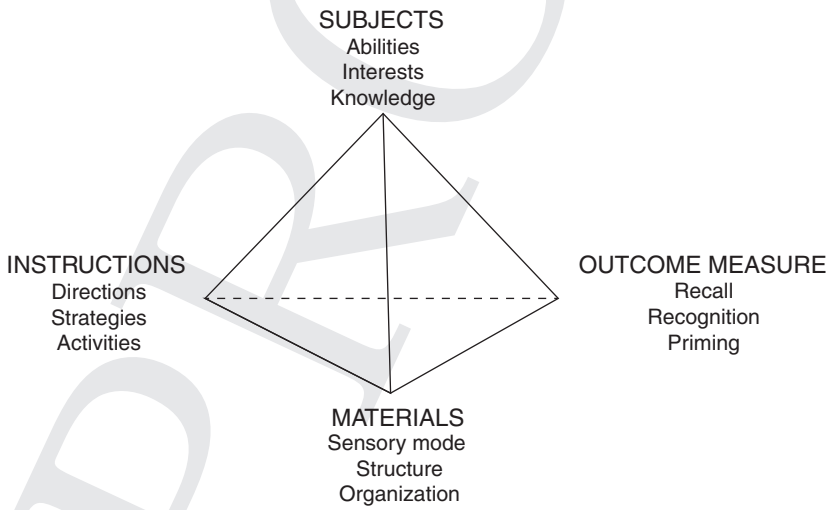

Figure 11.1 Jenkins's tetrahedral model of memory experiments
Adapted from J. J. Jenkins (1979)

older adults, people with schizophrenia, and so on. If a finding is obtained with samples of one type, will it generalize to other groups? Only future experimentation can say for sure. Similarly, on a different dimension, if some memory phenomenon is obtained by using lists of words, will the results hold when the *materials* are switched to prose passages, to pictures, to poetry, to scientific texts? Another critical issue is the *instructions* given to subjects in an experiment and the settings in which they are tested. What strategies do people use and how do the instructions and settings influence strategies? Might the results be affected if these were changed? Finally, there is the *outcome measure* itself, the dependent variable. Almost any psychological construct can be measured in at least several different ways. Would the same results be obtained if a different dependent measure had been selected?

All these questions are good ones, and no firm answers can be given in the abstract. Rather, further research examining these factors must be conducted to see over what variables the findings will generalize. As we already noted, discovering the boundary conditions for some experimental finding – discovering conditions under which the finding does not hold – is often critically important for understanding a phenomenon and developing an accurate theory about it. So, in considering any particular experimental finding, one should keep in mind the control variables that were *not* studied in the experiment. Many may turn out not to be critical and the result will generalize across them; however, others may be critical and their manipulation will show limitations of the observed result or that it will not generalize to other conditions. The basic idea of the contextual approach is to take a phenomenon and to try to affect it greatly by manipulating factors held constant in the original work. Only by doing this can researchers get a full picture of the phenomenon they are studying. Nearly every phenomenon will have some boundary conditions, that is, conditions under which it will not occur.

Henry L. Roediger, III: A Painful Example of Critical Thinking with Regard to Replication

The first author of this chapter was once caught on the horns of a dilemma, along with Kathleen McDermott, a colleague, as discussed in Roediger (2012). Briefly, we conducted an experiment in which we manipulated how students studied a list of words (either paying attention to surface features of the words or paying attention to their meaning, called a *levels-of-processing manipulation* following Craik and Tulving (1975)). This manipulation was

carried out within-subjects, so all subjects answered questions about surface features or meaning for half the words (counterbalancing tasks and words across subjects). The other variable in the experiment was the dependent measure; briefly, we measured recognition memory in two different ways, and the type of test was manipulated between subjects. One set of subjects received one type of recognition test instruction, and another group received the other one. We had a hypothesis about why one method should be better than the other, so we hoped to show a difference between the two methods. However, going into the experiment, we thought it unlikely that we would find one, but that it would be really interesting if we did.

To our delight and surprise, we found a statistically significant difference such that one recognition measure was superior to the other. Even more interestingly, the independent variable (the way the material had been studied) had a bigger effect on the recognition method that produced higher performance, so that measure seemed more sensitive than the other one. That is, meaningful processing of words was superior to processing of surface features on both recognition tests, but the effect was much larger for the measure of recognition that produced the best recognition. (You will see in a moment why we are not describing more fully the two ways we used to measure recognition memory.)

We showed our results to a few friends, one of whom was a visiting expert. Their advice: Write up the experiment as soon as possible and submit it to a good journal. We also submitted an abstract of the work to one of the primary conferences attended by experimental psychologists. But we were faced with a dilemma. Did we go with our one study, or should we try to replicate the study before submitting it? We dithered. The study involved many participants, because we were testing trainees at a nearby Air Force base in our research. The data seemed compelling; the results were statistically significant. All these considerations argued for publishing quickly. However, what if we were somehow wrong? What if the study did not replicate? It seemed unlikely, but the only sure way to convince ourselves was to try to replicate with a fresh sample of participants.

After agonizing over the decision, we decided not to submit until we had replicated the experiment using a different sample of subjects (with exactly the same materials and procedure). (During the replication attempt, we gave the scheduled talk at the meeting, saying that we were in the process of trying to replicate the results.) The second time we tested a reasonably large sample of university students using exactly the same design. To our surprise and dismay, neither of the original effects replicated! The two methods of testing recognition did not differ significantly in the replication, although there was a hint of an effect in the same direction as the one originally found. So now we had two experiments, one showing the effect and one that did not. The effect might still be real, but small (or fickle). The next step was to go back to the military base and conduct the same experiment on the

same computers with a new (and larger) sample of military recruits. Now the results were clear cut: There was no hint of any effect, no hint that the original result could be replicated. The effect just did not seem to be there. We were flummoxed.

Over the next several years, we tried several variations on our original "successful" experiment, trying the find a reliable procedure to obtain the effect. We never succeeded in obtaining the effect; the two recognition procedures seemed to produce identical results, completely undercutting our original hypothesis that we thought was so clever. We were disappointed in the outcome, but glad we ignored others' advice to publish and to replicate our work instead.

The critical thinking lesson: Always, always, replicate your own work before publishing, especially if you are claiming to find some new and unexpected result. Had we published the original finding, we would have sullied the scientific literature and sent other scientists on a wild goose chase in studying a nonexistent phenomenon. Lesson learned. Or perhaps we had already learned it, because we did choose to replicate before publishing – and we couldn't replicate our own effect.

Critical Questions to Ask about Experiments: A Summary

We conclude the chapter with a set of critical questions that you should ask about experimental research. Keeping these issues in mind while reading experimental reports will aid your critical analysis of the experimental literature. Similarly, keeping these questions in mind while designing your own experiments should help you become a better researcher, too.

1. What hypothesis is the research testing? Is it clearly stated?
2. Does the experiment follow from the assumptions in the hypothesis? Can you think of more effective methods to use in testing the hypotheses?
3. What are the independent variables being manipulated? Does the manipulation seem to be an effective one? Will it permit a fair test of the hypothesis?
4. Are other variables confounded with the independent variable that is being manipulated?
5. What dependent measures are being examined? Do they actually measure the construct of interest? Can you think of other measures that should be used?
6. What variables are being controlled? Are there others that should be controlled?

7. Did the author use a within-subjects or between-subjects design? Why do you think this choice was made? Is it the right choice?
8. If the author used a within-subjects design, were practice effects controlled by appropriate counterbalancing? Did the author consider differential carryover effects?
9. If a between-subjects design was used, did the author use an appropriate procedure (randomization, matching) for ensuring that the subject groups did not differ?
10. How would you rate the probable internal validity of the experiment? That is, do you think it can be directly replicated? Why?
11. What generality might the experiment have, what external validity? Do you think the same effects would be obtained with (a) different types of subjects; (b) different experimental settings or instructions; (c) different materials, examples, or procedures; and (d) different dependent measures?
12. Did the study have sufficient statistical power for all the tests reported?
13. Was the study pre-registered, and did analyses and findings conform with the pre-registered predictions? (This procedure still rarely occurs.)
14. If the study was pre-registered, were pre-registered hypotheses and analyses clearly distinguished from any exploratory analyses?
15. Is the uncertainty in the data adequately conveyed (e.g. with confidence intervals for effect sizes)?
16. Were researcher degrees of freedom adequately disclosed, and choices justified?
17. Are the data publicly available?
18. Does the experiment contribute to knowledge? What do you know after reading the study that you did not know before, and why does that knowledge matter?

Key Terms

Between-subjects design Experimental design where different groups of subjects are assigned to each level of the independent variable.

Boundary conditions The conditions under which an effect of interest will not generalize.

Conceptual replication Seeking the same general pattern of results as a prior experiment, but using methods different from the original procedure.

Confounding When some factor varies with, or is correlated with, the independent variable of interest, and possibly influences the dependent variable.

Contextualism Understanding of all the variables that were held constant for a particular experiment, but which could be manipulated (e.g. Subjects, Instructions, Outcome Measure, and Materials).

Control variable Factors that the experimenter could manipulate, but instead holds constant as much as possible.

Counterbalancing Balancing the order of within-subjects conditions between subjects, so as to reduce the impact of practice effects.

Dependent variable (DV) (or Outcome measure) Factor that is *measured* in an experiment; variation should depend on the level of the independent variable.

Differential carryover A special type of practice effect, where participation in one condition can greatly change performance in the other condition (but not necessarily vice versa).

Direct replication A re-performance of an experiment with all conditions and treatments as similar as possible to the original experiment.

Experimental materials The materials subjects learn, are exposed to, etc.

External validity Whether the findings of an experiment generalize to other settings.

Generalizability Whether an effect still holds under conditions other than those originally tested.

HARKing Hypothesizing After Results are Known; using statistically significant findings discovered through exploratory analyses to justify a hypothesis created after the fact.

Hypothesis Suggestion concerning the relation between two (or more) variables that exist in nature.

Independent variable (IV) A factor that is *manipulated* in an experiment.

Internal validity Whether the independent variable affected the dependent variable – Usually high in experiments.

Matched groups design Experimental design where subjects are assigned to groups such that some potential confound is equated between the groups.

***p*-hacking** Attempts to manipulate the statistical significance of a test (e.g. by stopping data collection once a test reaches statistical significance).

Practice effects When subjects participate in multiple within-subjects conditions, practice in one condition might affect performance in the next condition.

Random assignment In between-subjects experimental designs, subjects should be randomly assigned to each condition. This ensures that on average, members of the different conditions are equal except for experimental treatment.

Researcher degrees of freedom The set of choices scientists must make when collecting and analyzing data.

Subject variable An individual difference variable that cannot be randomly assigned, but is an intrinsic condition of subjects (e.g. age).

Subjects (or Participant) The people or (non-human) animals who are tested in an experiment.

Systematic replication Re-performance of an experiment, but manipulating new variables that, if an effect is robust, should not matter. Used to determine how robust an effect is, and to discover boundary conditions.

Variable Any factor in an experiment that can be manipulated or measured.

Within-subjects design Experimental design where the same individuals serve in all experimental conditions.

REFERENCES

Aarts, A. A., et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. DOI:10.1126/science.aac4716

Bacon, F. (1620/2000) *The new organon*. Eds. L. Jardine & M. Silverthorne. Originally published as *Novum Organum, sive indicia vera de Interpretationes Naturae* in 1620. Cambridge: Cambridge University Press.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist*, *44*, 1185–1193. DOI:10.1037/0003-066X.44.9.1185

Bem, D. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. DOI:10.1037/a0021524

Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition – A satire in one part. *Perspectives on Psychological Science*, *7*(3), 307–309. DOI:10.1177/1745691612441216

Boring, E. G. (1929). *A history of experimental psychology*. New York and London: Century.

Camerer, C. F., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature: Human Behavior*, *2*, 637–644. DOI:10.1038/s41562-018–0399–z

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 263–294. DOI:10.1037/0096–3445.104.3.268

Cummings, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Deese, J. (1959). On the prediction of occurrence of particular verbal instructions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22. DOI:10.1037/h0046671

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS ONE*, *7*(1), e2908. DOI:10.1371/journal.pone.0029081

Freyd, J. J., & Gleaves, D. H. (1996). "Remembering" words not presented in lists: Relevance to the current recovered/false memory controversy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 811–813. DOI:10.1037/0278–7393.22.3.811

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effect on girls' mathematics performance. *Developmental Psychology*, *49*(10), 1886–1897. DOI:10.1037/a0031412

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *37*, 833–848. DOI:10.3758/MC.38.7.833

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *26*, 1–104.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "*p*-hacking" and the research hypothesis was posited ahead of time. Online. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Hagger, M. S., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. DOI:10.1177/1745691616652873

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *P*-hacking in science. *PloS Biology*, *13*(3), e1002106. DOI:10.1371/journal.pbio.1002106

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PloS Medicine*, *2*(8), 0696–0701. DOI:10.1371/journal.pmed.0020124

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Hillsdale, NJ: Erlbaum.

Kahneman, D. (2012). A proposal to deal with questions about priming effects. Online. https://tinyurl.com/yyj8v7nx

Kantowitz, B. H., Roediger, III, H. L., & Elmes, D. (2015). *Experimental psychology* (10th ed.). Belmont, CA: Wadsworth.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. DOI:10.1207/s15327957pspr0203_4

Klein, R. A., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. DOI:10.1027/1864-9335/a000178

McDermott, K. B., & Naaz, F. (2014). Is recitation an effective tool for adult learners? *Journal of Applied Research in Memory and Cognition*, *3*, 207–213. DOI:10.1016/j.jarmac.2014.06.006

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387. DOI:10.1037/0003-066X.38.4.379

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. DOI:10.1177/1745691612465253

Piaget, J. (1962). *Play, dreams, and imitation in childhood*. New York: Norton.

Popper, K. (2002). *The logic of scientific discovery*. New York: Routledge.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, *26*(5), 653–656. DOI:10.1177/0956797614553946

Roediger, III, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *The APS Observer*, *25*(2), 9, 27–29.

Roediger, III, H. L., & Gallo, D. A. (2002). Processes affecting accuracy and distortion in memory: An overview. In M. L. Eisen, G. S. Goodman, & J. A. Quas (Eds.), *Memory and suggestibility in the forensic interview* (pp. 3–28). Mahwah, NJ: Erlbaum.

Roediger, III, H. L., & Gallo, D. A. (2016). Associative memory illusions. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (2nd ed.) (pp. 390–405). New York: Psychology Press.

Roediger, III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. DOI:10.1111/j.1467–9280.2006.01693.x

Roediger, III, H. L., & McCabe, D. P. (2007) Evaluating Experimental Research: Critical Issues. In R. J. Sternberg, H. L. Roediger, III, & D. F. Halpern (Eds.), *Critical thinking in psychology* (pp. 15–36). New York: Cambridge University Press.

Roediger, III, H. L., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814. DOI:10.1037/0278–7393.21.4.803

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. DOI:10.1177/0956797611417632

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. DOI:10.1177/1745691615609918

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 175–184. DOI:10.1016/S0022–5371(67)80092–6

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1–12. DOI:10.1037/h0080017

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, *70*, 122–129. DOI:10.1037/h0022014