Critical Thinking in Psychology

Edited by
ROBERT J. STERNBERG
Tufts University

HENRY L. ROEDIGER III
Washington University in St. Louis

DIANE F. HALPERN
Claremont McKenna College

CAMBRIDGE
UNIVERSITY PRESS

2

Evaluating Experimental Research

Critical Issues

Henry L. Roediger III and David P. McCabe

> [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.

The author of this quote is Edwin G. Boring (1886–1968), one of the great psychologists of the 20th century and author of *A History of Experimental Psychology* (1929; the quote comes from 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 150 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 thoughts from the ancient Greek, Roman, and Arab civilizations were rediscovered. The 17th 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 humans or their mental life. The first physiologists and anatomists 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 occupy 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 to 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 (published in *Novum Organum* in 1620, as translated in 2000) 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" (p. 143). 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, 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). Critical thinking – converting hypotheses into experimental tests – is at the heart of experimental methods.

THE EXPERIMENTAL METHOD

The heart of the experimental method is straightforward. A theory or hypothesis 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. 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 the recall of the passages 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, confounded 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 (type 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 (of 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 whether to use 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 20 times and a different group of subjects would study the passage 10 times and be tested 10 times in 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.

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 20 times and then, after doing that, they would study a second passage 10 times with 10 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 confounding 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, 2005).

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. Yet, as just discussed, there are drawbacks to within-subjects designs, too, because the same people are tested in each condition, and serving in one condition can affect performance in the other condition. If a differential carryover effect is likely to occur and to cloud the results, 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 greatly influence the other conditions in the experiment, 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. When individual differences among people are examined on some task or set of tasks, the factor is referred to as a *subject variable.* 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 differences 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, age in this case.

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 for the remainder of the chapter. 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 Elysees, 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 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 15 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 15 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 to 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 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 (e.g., Roediger, Balota, & Watson, 2001; McDermott & Watson, 2001).

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 at a time for 1.5 seconds per word. After 8 lists they recalled as many words as they could for 2 minutes, whereas after the other 8 lists they did arithmetic problems 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. The subjects were cued after each list as to whether they would do arithmetic or recall the list, so they probably listened to the lists expecting to recall them in all cases. They

never knew until after the list was over whether or not they would have to recall it.

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 recognition 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 nonpresented words (called lures or distracters 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 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 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% of the list words. However, they also recalled the critical (nonstudied) items from those lists 55% 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 subjects were instructed to be sure to write down only words that they had just heard.

The recognition test results are provided in Table 2.1. At the top half are results for list words (from lists studied and recalled, from those studied but not recalled, and from those that were not studied at all). The proportions given under the Overall column are for the proportion of items judged to be "old." This number is then decomposed into those old responses that were judged to be remembered and those judged to be known in last two columns. The data in the top half of the table show that prior recall boosted later recognition (the effect of testing that we have already discussed) and that this effect occurred because of increases in remembering (Know judgments for the two conditions were about equal). Notice that the subjects were much more likely to say that they *remembered* than *knew* the words if the lists had been studied. For the nonstudied lists, the false alarm rate (probability of calling an item studied) was quite low (.11), and in this case almost all the

TABLE 2.1. *Recognition Results for the Roediger–McDermott Experiment*

Item Type and Condition	Proportion of "Old" Responses		
	Overall	Remember	Know
List Words			
Study + Recall	.79	.57	.22
Study + Arithmetic	.65	.41	.24
Nonstudied	.11	.02	.09
Critical Lures			
Study + Recall	.81	.58	.23
Study + Arithmetic	.72	.38	.34
Nonstudied	.16	.03	.13

Source: From Roediger and McDermott (1995).

responses are said to be *known* and not *remembered.* This last finding makes sense: How could someone remember an event that never happened?

The answer to this question is in the bottom half of Table 2.1: If a person has experienced events that are associated to (and strongly imply) another event, they may remember the event as actually having occurred. The data for the critical items showed that people falsely recognized them at high levels for the studied lists (.81 if the list had been recalled and .72 if it had not), but if the list had not been studied the false alarm rate was low (.16). Even more surprisingly, people not only recognized the nonpresented words as having been heard, but they judged that they remembered the moment of occurrence of the words at the same levels as for the presented words! That is, the *Remember* judgments for the critical (nonpresented) items in the bottom half of Table 2.1 are of the same magnitude as for the presented items at the top, so this outcome truly represents false remembering.

The Roediger and McDermott (1995) results show that even in the simple task of recalling a list of words, people can suffer from false recollections. These results have been studied and debated for the past 11 years, so we use this paper as a target to critically evaluate experimental research.

GENERALITY AND LIMITATIONS OF EXPERIMENTAL RESULTS

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, p. 812) 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.

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. We hazard the guess that most experimental results in psychology can be directly replicated (although there are some notable exceptions to this claim). A conceptual replication is the next step. A *conceptual replication* refers to seeking the same general pattern of results but using somewhat 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 slightly differently 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 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.

JENKINS'S APPROACH TO 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 should be considered in the context of the factors that were held constant but that could have potentially been varied. That is, that every experiment occurs in the context of control variables or factors that were held constant in the research, but that could have been

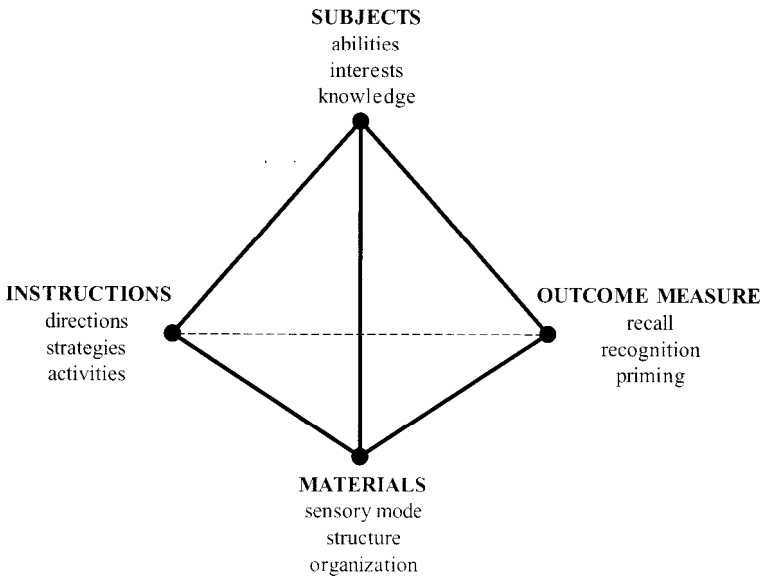

FIGURE 2.1. Jenkins' tetrahedral model of memory experiments. Adapted from J. J. Jenkins (1979).

manipulated. To what extent is the phenomenon of interest determined by the setting of the control variables? Posing the question a different way, when other variables are manipulated, will the experimental effect still generalize to these conditions? This is one way of thinking about the issue of conceptual replications.

Jenkins' basic idea of *contextualism* is represented in Figure 2.1, 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, 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. What strategies do people use and how do the instructions influence strategies? Might the results be affected if these were changed? Finally, there is the *outcome measure* itself, the dependent variable. Most 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 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 to the observed result. We will now consider how each of the four factors that Jenkins mentioned – subjects, materials, instructions, and outcome measures – affects false memories in the DRM converging associates paradigm as an example of how researchers have critically approached this topic.

Subjects

Different groups of subjects often vary considerably in their level of memory performance, and subject variables turn out to be particularly important in influencing false memories. Some subjects, like children and people with amnesia, typically show a smaller proportion of false memories than do the college students who were originally used in the DRM paradigm (see Brainerd & Reyna, 2005, chapter 5). Other subject groups, such as older adults and people with Alzheimer's disease, typically show more false memories compared with college students (e.g., Balota et al., 1999).

As mentioned earlier, the converging associates paradigm has been criticized as unnatural, and unlikely to generalize to the real world (Freyd & Gleaves, 1996), but of course this is an empirical question and one could argue the converse point, too (Roediger & McDermott, 1996). One approach to verifying that similar processes are important in the laboratory and the real world of memory distortion is to examine whether people who are susceptible to memory errors in the real world are also more susceptible to false memories in the laboratory. One particularly interesting group of people that fit the criteria of being susceptible to memory distortions is people who claim to have been abducted by aliens and to remember their experiences.

Clancy, McNally, Schacter, Lenzenweger, and Pitman (2002) examined whether alien abductees – people who believe they have been abducted by aliens – were more likely to falsely remember related lures in the DRM converging associates paradigm, as compared with people who did not claim to be abducted. The alien abductees and control subjects were recruited by using a newspaper advertisement, and were matched on age and education. Subjects were brought into the lab and participated in an experiment

similar to that of Roediger and McDermott (1995). In the Clancy et al. study, subjects studied lists of 3, 6, 9, 12, or 15 associates, each of which was related to a lure that was not presented. Previous research had shown that false recall increases with increasing list length (Robinson & Roediger, 1997), so this new experiment represented a direct replication of that previous study. After each list, subjects completed math problems for 30 seconds and then tried to recall the words they had heard. Following study and recall of all lists, subjects took a recognition test for all the lists. Because the number of words in a list did not interact with the subject variable, the results will be presented as an overall average across all list lengths. The primary finding of interest was that alien abductees were more than twice as likely as control subjects to falsely recall related lures (.29 vs. .14, respectively). A similar result was found for false recognition, with abductees falsely claiming to recognize related lures more often (.67) than did control subjects (.42). Because there were no differences between the groups in terms of their memories for the actual studied words or for the other distracters on the recognition test, it appears that the alien abductees were more likely to have difficulty differentiating words that were studied from those that were only implicitly activated. Of course, as mentioned previously, when examining subject groups who differ on some important dimension like their likelihood of reporting false memories in the real world, we should be careful to examine what other ways these subjects differ as well. For example, the alien abductees were also more likely than controls to report feeling depressed and exhibiting other symptoms of mental illness, and these reports were positively related to the likelihood that they would exhibit false memories. Thus, it is unclear if depression causes increases in false memories, or whether it is actually a difference in the memory processes of alien abductees that causes this difference. Nonetheless, this study verifies that people who likely have experienced false memories in the real world are also more susceptible to false memories in the laboratory, which suggests that the memory processes involved in false memories in the laboratory are similar to those involved in false memories in the real world.

Outcome Measures

In discussing the Clancy et al. study on false memories in the previous section, we mentioned that the results were similar for both false recall and false recognition. This is an example of a result generalizing across another critical factor mentioned in Jenkins' contextual model of memory: the outcome measure or the dependent variable. Memory research, like other areas of experimentation, has different types of outcome measures that are sensitive to different types of manipulations. For example, one ubiquitous finding in memory research is that words that are used less frequently in the English language are identified better on recognition tests than words that are used

more frequently but are not recalled better. These lower frequency items are likely better recognized because they are more distinct than higher frequency items, making them "stand out" on recognition tests. By contrast, on recall tests this type of distinctiveness is apparently not as beneficial to memory, so these lower frequency words are not better recalled. Thus, it is often the case that different memory tests need to be used to generalize results across a variety of measures. With respect to false memories, does the type of memory test matter?

For the most part, research indicates that both measures of recall and recognition show similar patterns of results in most false memory experiments; manipulation of an independent variable or subject variable usually affects false recall in the same manner as it affects false recognition (as in the study by Clancy et al., 2002). For example, false recall and false recognition both decrease after a delay. However, if people must think about the meaning of the words while studying lists rather than thinking about letter characteristics (e.g., number of vowels in the word), both measures of false memory (recall and recognition) increase (Thapar & McDermott, 2001). Also, repeating studied words several times decreases false recall and recognition (Kensinger & Schacter, 1999), as does having subjects generate the words from anagrams (McCabe & Smith, in press). Finally, aging and amnesia usually affect false recall and false recognition in similar ways (Kensinger & Schacter, 1999; Schacter, Verfaellie, & Pradere, 1996). Thus, although recall and recognition tests sometimes lead to different results depending on the nature of the independent variables manipulated in an experiment, this is not typically the case with false memories.

Materials

Jenkins also raised the issue of whether an effect would generalize across varying materials and methods of presentation. Would the DRM false memory effect generalize when different materials, modalities (visual vs. auditory), and organization were used to present the materials? The converging associates illusion has actually been found to be unusually robust, and generalizes across most conditions that have been studied, but the strength of the effect is influenced by the specific materials used and their organization. For example, in the original study (Roediger & McDermott, 1995), all the words associated with a particular related lure were studied together. Previous research using words from different categories shows that presenting words from the same category together enhances memory for those words relative to a condition in which words from several categories are mixed together. This is probably because the organization of the lists in a blocked format allows subjects to relate the words to one another, which provides effective retrieval cues on a memory test. Therefore, this "blocked" presentation of the studied words in the original Roediger and McDermott (1995)

study likely supported good memory for the studied words, but did it affect false memories?

McDermott (1996) first showed that blocked presentation increased both correct and false recall relative to random presentation. We discuss a follow-up experiment by Mather, Henkel, and Johnson (1997) that conceptually replicated the Roediger and McDermott (1995) and McDermott studies with a subset of their lists and a recognition memory procedure. Their design was a bit complicated, so we restrict our discussion to the simplest condition. In one condition, 10 lists were used and the words associated to a particular nonstudied word were blocked together. The 10 lists were studied successively, and this was followed by a final recognition test. In the other condition, studied words associated with all 10 related lures were randomly mixed together on the study list. Mather et al. (1997) found, as expected, that the identification of studied words on the recognition test was better when the words were blocked (.84) than it was when they were randomly mixed together (.69). However, blocking the studied words also led to more false recognition (.86) than when the words were randomly intermixed (.66). Thus, the studies by McDermott (1996) and by Mather et al.(1997) show that the organization of materials can affect both false recall and false recognition in the same manner, and these effects parallel the effect on accurate recognition.

Interestingly, this outcome is not always obtained when other independent variables are manipulated. Some variables enhance accurate recall or recognition while reducing false recall and false recognition. For example, reductions in false memories are found when studied words are presented visually as opposed to auditorily, or when studied words are presented in distinctive fonts, or as anagrams (see McCabe, Presmanes, Robertson, & Smith, 2004 for a review). In each of these cases, making the studied words more distinctive appears to lead to reduced levels of false memories and increased levels of accurate retention.

Instructions and the Experimental Setting

Another critical issue in determining generality is, for lack of a better term, the general setting of the experiment – the instructions used, the particular experimenter, the room where it is conducted, the expectations the subjects are given, and the strategies they use. Some of these factors do turn out to be important in determining how often subjects will falsely remember related lures. In one example, different groups of subjects were trained to use one of two different encoding strategies while they were studying the words (McCabe et al., 2004). The question of interest was whether increasing the distinctiveness of the studied words would decrease the likelihood that people would falsely recognize related lures. In the item-specific rehearsal condition, which was intended to encourage distinctiveness, subjects were

told to think about one unique characteristic of each studied word, some feature that made each word different from the other words in that list. The other group, the relational encoding group, was told to think about what the words had in common. The hypothesis was that the item-specific encoding group would be better able to distinguish the studied words from each other compared with the relational group, and would also be better able to distinguish the related lures from the studied words. The results were consistent with the hypothesis that making studied words more distinctive led to reductions in false recognition. Related lures were falsely endorsed only 64% of the time in the item-specific encoding group, but 84% of the time following relational encoding. This outcome was replicated in a within-subjects design, when half of the lists were studied under either encoding strategy, indicating that the results were not due to strategy differences between the two groups on the tests, but rather were the result of differences in memory for the studied words themselves. This outcome (and others like it) shows that false recall and false recognition in the DRM converging associates paradigm can be reduced (although not eliminated) by using different encoding strategies created by different instructional sets (McCabe & Smith, in press).

Summary of Jenkins' Contextual Approach

The examination of false memory research from the Jenkins perspective was revealing with respect to how these effects generalize. First, we can conclude that the converging associates memory illusion is quite robust, having never been eliminated across the different subjects, outcome measures, materials, or instructions we considered. Second, we found that false memories differed across subject groups, types of materials, and instructions, but they were relatively unchanged by the type of outcome measure used to assess memory performance. On the basis of the data reviewed we can conclude that false memories are difficult to eliminate, but they can be reduced when memory for the studied items is distinct and subjects are willing and able to discriminate what is real from what is not. The important point to note is that these conclusions can only be made after a careful examination of the contextual factors that can affect memory performance. The basic idea of the contextual approach is to take a phenomenon and try to greatly affect it 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.

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? 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; (c) different materials, examples, or procedures; and (d) different dependent measures?
12. Does the experiment make a contribution to knowledge?

References

Bacon, F. The new organon. (L. Jardine & M. Silverthorne, trans.). Cambridge: Cambridge University Press. (Original work published in 1620).

Balota, D. A., Cortese, M. J., Duchek, J. M., Adams, D. A., Roediger, H. L., McDermott, K. B., & Yerys, B. E. (1999). Veridical and false memories in healthy older adults and in dementia of the Alzheimers type. *Cognitive Neuropsychology, 16*, 361–384.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist, 44*, 1185–1193.

Boring, E. G. (1929). *History of experimental psychology.* London: Century/Random House.

Brainerd, C. J. & Reyna, V. F. (2005). *The science of false memory.* Oxford: Oxford University Press.

Clancy, S. A., McNally, R. J., Schacter, D. L., Lenzenweger, M. F., & Pitman, R. K. (2002). Memory distortion in people reporting abduction by aliens. *Journal of Abnormal Psychology, 111,* 455–461.

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

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.

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

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.

Kantowitz, B. H., Roediger, H. L., & Elmes, D. (2005). *Experimental psychology* (8th ed.). Belmont, CA: Wadsworth.

Kensinger E. A., & Schacter D. L. (1999). When true memories suppress false memories: Effects of aging. *Cognitive Neuropsychology, 16,* 399–415.

Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know and memory characteristics questionnaire compared. *Memory & Cognition, 25,* 826–837.

McCabe, D. P., Presmanes, A. G., Robertson, C. L., & Smith, A. D. (2004). Item-specific processing reduces false memories. *Psychonomic Bulletin & Review, 11,* 1074–1079.

McCabe, D. P., & Smith, A. D. (in press). The distinctiveness heuristic in false recognition and false recall. *Memory.*

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language, 35,* 212–230.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory and Language, 45,* 160–176.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

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

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science, 8,* 231–237.

Roediger, H. L., Balota, D. A., & Watson, J. M. (2001). Spreading activation and the arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association.

Roediger, 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, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255.

Roediger, 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.

Roediger, H. L., & McDermott, K. B. (1996). False perceptions of false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 814–816.

Schacter, D. L., Verfaellie, M., & Pradere, D. (1996). The neuropsychology of memory illusions: False recall and recognition in amnesic patients. *Journal of Memory and Language, 35*, 319–334.

Thapar, A., & McDermott, K. B. (2001). False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. *Memory & Cognition, 29*, 424–432.

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.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*, 1–12.

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