3

Internal and External Validity in Clinical Research

Steven Taylor

Gordon J. G. Asmundson

Advances in clinical psychology critically depend on methods researchers use for investigating the causal relationships among variables. Research questions commonly include issues about the putative mechanisms associated with various forms of psychopathology (e.g., do dysfunctional beliefs cause depression?) and questions about the effects of treatments (e.g., does cognitive therapy cause a greater reduction in eating disorder symptoms than does placebo?). How do we go about drawing objective conclusions about causality? This is a question about whether the manipulation of one variable (i.e., the independent variable) has effects on another variable (i.e., the dependent variable). The answer to this question is more complex than it might seem. Consider the following common scenarios.

Scenario 1: *Dr. Smith is a well-known local advocate of a controversial form of psychotherapy. He claims that it works faster and more powerfully than all other treatments. For many years he has practiced this treatment and trained other clinicians through workshops, even though there were no scientific data on its efficacy. When challenged on this point, Dr. Smith retorted that he had "seen all the proof he needed with his own eyes." That is, he claimed his patients almost invariably benefited from his therapy. When a treatment outcome study was conducted and published by an independent group of investigators, it was found that the psychotherapy used by Dr. Smith was no more effective than giving patients a placebo. Dr. Smith's response was "there must have been something wrong with the research; the researchers didn't include patients seen in the real world."*

Scenario 2: *The interns and their clinical supervisors gathered in the seminar room for the weekly journal club. This week's article, which recently appeared in a leading journal, described an experimental investigation of the effects of cognitive factors (expectations of disapproval) on social anxiety. Research participants led to expect high disapproval (from an experimental confederate) experienced more social anxiety than participants who were led to expect no disapproval. The investigators concluded that expectations of disapproval can cause social anxiety and likely play a role in clinically severe conditions such as social anxiety disorder. Despite the many methodological strengths of the study, the participants of the journal club quickly searched out the methodological weaknesses. One of the interns raised a particularly important*

question about the generalizability of the study. Her observation was met with nods of approval from her supervisors. By the time the journal club had finished, all the attendees had convinced themselves that the study was fatally flawed. Some attendees even left the meeting with the impression that psychological research, even research published in leading journals, is largely a waste of time.

This chapter is written largely in response to these two types of scenarios, which we have encountered time and again. The reactions illustrated in these scenarios retard the scientific progress of clinical psychology. The first scenario raises many issues regarding internal and external validity. Dr. Smith claims that all his patients benefit from his treatment. Yet, his series of case observations have many problems of internal validity. His dismissal of a recent study of his psychotherapy raises the issue of external validity. The journal club scenario raises the question of what we can conclude from research that does not have "perfect" internal and external validity.

The important issues raised in these scenarios are the focus of the remainder of this chapter. We will begin by defining internal validity and illustrating the various threats to it. Some studies, such as those using quasi-experimental designs, are widely used in clinical research, despite their imperfect internal validity. After reading this chapter, you should have a good understanding of why such studies are used and why they are useful. After discussing internal validity, we then examine the concept of external validity (generalizability) and consider the relationship between internal and external validity. As you will see, there are some research situations in which high external validity is vital and other situations in which it is not a priority. Finally, we will conclude with some comments about how scientific knowledge can be advanced even though most research studies have imperfect internal or external validity. Throughout this discussion we will consider a number of commonly used experimental designs that have been developed to deal with issues of internal and external validity. Our discussion of these designs will be illustrative rather than comprehensive. Detailed discussions of experimental and quasi-experimental designs are available elsewhere (e.g., Asmundson, Norton, & Stein, 2002; Barlow & Hersen, 1984; Campbell & Stanley, 1970; Cook & Campbell, 1979; Onghena & Edgington, 2005).

Internal Validity

Internal validity is the degree to which observed changes in a dependent variable can be attributed to changes in an independent variable. Thus, internal validity is a matter of degree (e.g., high, medium, low) rather than one of presence or absence. The researcher's confidence in his or her findings is proportionate to the strength of internal validity of the research design (Finger & Rand, 2003). True experiments are designs that have strong internal validity; that is, participants are randomized to experimental conditions, and other means are used to ensure that changes in the dependent variable can be attributed to the experimental manipulation of the independent variable. Quasi-experimental designs have weaker internal validity, as we will illustrate later. There are several types of threat to internal validity (Cook & Campbell, 1979; Finger & Rand, 2003; Rosenthal, 2002), including

- history,
- maturation,
- testing,
- instrumentation,
- statistical regression,
- attrition,
- selection,
- interactions with selection,
- diffusion or imitation of treatments,
- compensatory equalization of treatments, and
- experimenter expectancy.

Each of these threats to internal validity are defined and illustrated in the following sections.

History

Description. When changes in the dependent variable are due to some extraneous event that takes place between pre- and posttest, it makes it difficult to determine whether the results were due to the experimental manipulation (i.e., changes in the independent variable) or to the extraneous event. In some research, such as a short study of memory, this threat can be controlled by shielding participants from outside influences during the study (e.g., testing them in a quiet lab) or by choosing dependent variables that could not plausibly have been affected by

outside influences (Cook & Campbell, 1979). In treatment studies, which may take the participant several weeks to complete, these methods may not effectively control for the effects of history. Other methods are more often used, such as random assignment of participants to an experimental group or a control group. In a therapy study, the latter might be a waiting list control in which the participants are simply assessed twice, with the retest interval being matched to the duration of the treatment study. Participants in the treatment condition would be similarly assessed twice, before and after treatment.

Example. Midway through an uncontrolled study of treatments of driving phobia, a well-known celebrity was killed in a car accident, thereby inflating the fears of the treatment participants. As a result, their posttreatment scores on a measure of driving fear tended to be higher than their pretreatment scores, thereby giving the misleading impression that treatment worsened their phobias. Inclusion of a waiting list control group would have demonstrated the impact of this event on those with driving phobia who were not receiving treatment.

Maturation

Description. Change in the participant over the course of time, where such change is not the focus of interest of the research study. This may involve growth (e.g., getting smarter or stronger) or decline (e.g., dementing). This threat can be addressed by using a control group.

Example. A drug company treated 20 elderly people in the early stages of Alzheimer's disease with a new medication for depression. The investigators concluded that the drug was effective in alleviating depression in this population. However, they failed to realize that depression naturally remitted for many patients as their dementia worsened. That is, as their memories became worse, the participants no longer had insight into the fact that they were dementing and so were no longer depressed about this problem. Inclusion of a control group (in this case, one receiving a placebo pill) would have allowed the researchers to assess natural changes in depressive symptoms associated with increasing dementia.

Testing

Description. The reactive effects of testing where the very act of assessment influences the variable under investigation. Some measures are highly reactive, whereas other measures are largely unreactive. Also, repeated testing can increase familiarity with the test, which might bias scores. This threat can be dealt with in various ways, such as by selecting unreactive measures (e.g., unobtrusive observation) or by including a control group.

Example. In studies of smoking, the act of monitoring one's use of cigarettes affects the frequency of smoking. That is, self-monitoring may help some people realize how much they smoke, thereby motivating them to cut down. To illustrate the effects of test familiarity, consider a study in which tests of intelligence are administered on multiple occasions. With repeated testing, participants may become better at some tests (e.g., the digit-symbol subtest of the Wechsler Adult Intelligence Scale) simply because they have learned the correct responses (e.g., the symbol that goes with each digit) as a result of repeated testing.

Instrumentation

Description. When an effect is due to a change in the measuring instrument from pre- to posttest rather than due to the manipulation of the independent variable. Instrumentation can affect all forms of measurement, including observers, self-report tools, interview schedules, and devices that measure physiological processes.

Example. In observational studies, progressive fatigue of observers who are coding various types of marital interaction can impact their rating accuracy. With increasing fatigue, the observers may be less likely to detect subtle interaction patterns. Using one scale at pretest (e.g., the first edition of the Beck Depression Inventory) and another edition at posttest (e.g., the second edition of this inventory) might suggest a change in depressive symptoms where there was no change. Similarly, in a study measuring physiological reactivity to stress before and after stress inoculation training, changes in equipment calibration might falsely indicate or mask a treatment effect.

Statistical Regression

Description. People selected for extreme scores (very high or very low) will have less extreme scores when they are retested on the same or related variables. Why does regression occur? The

farther a score is from the mean, the more extreme it is. The more extreme the score, the rarer it is and the more likely it is to have been the result of a very rare combination of factors. If these factors are temporally unstable, then statistical regression will occur (Furby, 1973). Statistical regression is always toward the population mean of the group. Its magnitude is greater when the test-retest reliability of a measure is low (indicating that scores are readily influenced by chance factors) and when a person's score is extreme, relative to the mean of the population from which the person was chosen (Cook & Campbell, 1979). Regression effects will not be a threat if assessment methods are chosen that are virtually error free or uninfluenced by random factors (e.g., measuring a person's height; Finger & Rand, 2003). It is important to note that statistical regression effects can be due to psychologically substantive phenomena and should not be automatically dismissed out of hand as statistical artifacts (Taylor, 1994). Regression might be either noise or the phenomena of interest, depending on one's research goals.

Examples. A gambling researcher screened a large group of students in order to identify people who could be classified as heavy gamblers, as measured by a questionnaire. When these people presented to the lab to participate in the experiment, they completed the questionnaire a second time. To the researcher's chagrin, many of the participants no longer had extreme scores on the questionnaire and had to be excluded from the study. Another example concerns uncontrolled treatment studies, in which a group of patients are selected on the basis of extreme scores on some measure (e.g., scores on a perfectionism scale), and then receive an intervention (e.g., a treatment for excessive perfectionism) as well as a posttest. Statistical regression may occur, resulting in what appears to be a treatment effect (i.e., a decline in scores from pre- to posttest). The solution to this problem is to include a control group. Note that statistical regression is unlikely to occur if participants are selected because they have persistently elevated scores, such as people with chronically high scores on a measure of anxiety (i.e., high trait anxiety). Such people are unlikely to show statistical regression because this phenomenon is due to transient factors that produce elevations in scores (e.g., a near-miss while driving to the lab to participate in an experiment would transiently increase one's anxiety).

Attrition

Description. The loss of participants from a study (e.g., due to mortality or treatment dropout). This threatens internal validity if attrition is not random: for instance, if attrition is greater in one experimental condition than another, or if particular participants are most likely to drop out of the study. Attrition can cause serious problems for clinical researchers by introducing biases into an experiment. There are various methodological and statistical procedures for limiting, evaluating, and correcting for attrition (see Flick, 1988). However, there are circumstances in which attrition can render the results uninterpretable, as illustrated in the following example.

Example. A residential treatment center reported that 80% of patients with anorexia nervosa were "much improved" or "greatly improved" after completing the program. Unfortunately, the results were biased because 20% of patients did not complete the program, and no outcome data were available for them. Some withdrew because they benefited quickly from treatment and felt that they no longer needed to be in the program. Others dropped out because they failed to benefit. Some severely anorexic patients had either died or withdrawn from the clinic and were admitted to hospital. Given the large proportion of treatment dropouts and the uncertainty about whether treatment completers differed, as a group, from treatment dropouts, it was not possible to draw any legitimate conclusions from the treatment study.

Selection

Description. When the effects on the dependent variable arise from differences in the kinds of people in the experimental groups. Selection effects are pervasive in quasi-experimental designs (Cook & Campbell, 1979). These are among the most widely used designs in clinical psychology, in which a target group is compared with one or more control groups (e.g., a group of healthy people, a group with another psychopathology). Attempts are made to match the groups on background variables (e.g., demographics), and then they are compared on the variables of interest. However, assignment of participants to groups (e.g., target group vs. healthy control group) is, by definition, nonrandom in quasi-experimental designs. One must remember

that the distinction between the target group and any control group is an *observed* not experimentally manipulated distinction.

Example. Many studies have investigated whether people with anxiety disorders, as compared with healthy controls, tend to selectively focus their attention on sources of threat in their environment (Mogg & Bradley, 1999). Although such studies have yielded a good deal of useful information and have stimulated a great deal of research, these studies are prone to selection effects. That is, although the clinical (target) and control groups were matched on many background variables, there is no guarantee that the differences between the groups (e.g., the threat-focused attention effects) were due to the presence versus absence of an anxiety disorder. The effects could have been due to other factors that were not assessed in the study. In these studies, selection effects are addressed in three ways. First, the plausibility of confounding factors is taken into consideration. Anxious patients and healthy controls could differ on an almost infinite range of factors. Some of these factors could confound the study of threat-focused attention (e.g., depression), while other factors are less plausible (e.g., the participant's Zodiac sign). Second, researchers try to control for all the plausible confounding factors (e.g., all participants are asked to refrain from caffeine consumption on the day of testing; testing is done under conditions of normal or corrected-to-normal vision). Finally, if another confounding factor is subsequently identified (e.g., whether or not the person is taking antianxiety medication), then the study can be replicated, controlling for this factor.

Interactions With Selection. Many of the previously mentioned threats to internal validity can interact with selection to produce effects on the dependent variable that may be confused with effects due to the independent variable (Cook & Campbell, 1979). Examples include selection-history, selection-maturation, and selection-instrumentation effects. Selection-maturation interactions occur when the experimental groups mature at different speeds. Selection-history interactive effects occur when different experimental groups come from different settings, where each setting is associated with different histories. In a study designed to test the hypothesis that people with hypertension tend to be high in trait anger (i.e., anger proneness), for example, the hypertensive patients might tend to live in stressful environments, whereas normotensive controls might tend to live in relatively low stress environments. Thus, the different histories of the groups (i.e., differences in environmental stressors) might be responsible for any group differences in anger proneness, even if anger is unrelated to hypertension.

Diffusion or Imitation of Treatments

Description. When participants in the different experimental conditions can communicate with one another, such that participants in one condition learn about what happens in the other condition. This can undermine the differences between the experimental manipulations in each condition.

Example. In a study of the effects of stress (in the form of electric shock) on snake phobias, snake-fearful undergraduate psychology students were randomly assigned to one of two groups. Participants in each group were tested individually. Each participant was asked to walk up to a container housing a large, harmless snake and to touch it. Participants in the experimental group received a painful electric shock at a randomly determined point as they approached the snake. Participants in the control group experienced no shock as they approached the snake. Unfortunately, the students who had completed the experiment described their experiences with students who were soon to participate in the study. This contaminated the experimental manipulation because many of the people in the control group had heard about the electric shock. As they approached the snake they worried about getting a shock. Thus, the "no shock" control condition was compromised. Possible solutions to the problem of diffusion of treatments is to ask participants not to discuss the experiment with other students (during the period in which the study is being conducted) or to use experimental designs in which diffusion is not an issue (e.g., in the snake example, one could explicitly inform the control participants that they have been allocated to a "no shock" condition).

Compensatory Equalization of Treatments

Description. When participants learn that they have been assigned to an experimental condition where they won't receive the possible benefits received by participants in another experimental

condition. Participants may be reluctant to tolerate this inequality and thereby seek out the potential benefits.

Examples. In a study examining the effects of biofeedback for tension headaches, participants were randomly assigned to either biofeedback or a no-treatment (waiting list) control. Participants in the control group were aware that participants in the other group were receiving a potentially beneficial treatment. This perceived inequality prompted some people from the control group to seek out headache treatment while they were in the waiting list condition. This confounded the investigation of the effects of biofeedback. Another example concerns the use of pill placebo in drug studies. Participants in these studies are informed that they will be randomly assigned to receive either capsules containing the drug under investigation or capsules containing an inert substance (placebo). Unfortunately, in many drug studies, it is not difficult for patients to discover whether they have been assigned to the drug or placebo conditions, because drugs, unlike placebos, commonly produce side effects. Antidepressant medications, for example, may produce dry mouth or temporary jitteriness as side effects. Participants who discover that they are taking placeboes may therefore seek out additional treatments during the experiment, thereby confounding the investigation of the effects of the drug. Alternatively, some participants who realize that they are taking a placebo may become further depressed about not getting the "real" treatment. A solution to such confounds is to use placeboes that produce side effects. Some studies have taken this approach (Margraf et al., 1991).

Experimenter Expectancy

Description. A phenomenon whereby the participant's responses are influenced by expectations of the experimenter (or a proxy for the experimenter, such as a therapist or research assistant conducting a component of a study; Rosenthal, 2002). In other words, the participant's responses are shaped in the direction of the experimenter's expectations. These effects may be unintentional on the part of the experimenter (or therapist). This bias, sometimes known as the "allegiance effect," can be circumvented by keeping the people running the experiment (e.g., therapists, research assistants) blind to the aims of the investigation and by using procedures to counter any expectation effects of therapists.

Examples. In studies of a novel treatment, compared with some standard treatment, therapists may be highly enthusiastic about the new treatment and less impassioned by the standard treatment. A similar problem was encountered in our recent randomized, controlled study of three treatments for post-traumatic stress disorder (PTSD): behavior therapy, relaxation training, and eye movement desensitization and reprocessing (EMDR, Taylor et al., 2003). Some therapists were enthusiastic advocates of behavior therapy, while others were equally enthusiastic about EMDR. To control for possible expectancy effects, we had two therapists deliver all three treatments. One therapist was an expert and advocate of EMDR, while the other had expertise in behavior therapy. Thus, the therapists had potentially opposite expectations. This design enabled us to assess whether these and other therapist factors influenced treatment outcome. In this study, the treatments differed in efficacy (behavior therapy tended to be most effective), whereas there were no differences in the efficacy of the therapists, and there was no treatment-by-therapist interaction.

Comment

Returning to the first scenario that opened this chapter, we can see that most of these threats to internal validity would apply to Dr. Smith's observations about the effects of his treatment. His conclusions that his treatment was highly effective were based on simple pre/post case studies, that is, on patients assessed before and after this therapy. These studies failed to control for history, maturation, and statistical regression. Attrition was also a problem. A number of patients started treatment with Dr. Smith but dropped out because they failed to benefit from therapy. Dr. Smith conveniently failed to include these patients in his appraisal of his treatment's efficacy.

Not all case studies suffer as much from threats to internal validity. Various single-case experimental designs (which are part of the family of quasi-experimental designs) have been developed to deal with these threats (e.g., Barlow & Hersen, 1984; Onghena & Edgington, 2005). To illustrate, as part of our research into

cognitive-behavioral therapy (CBT) of panic disorder, we conducted a case study of an unusual presentation, in which the patient's panic disorder appeared to arise from blood-injury reactivity (vasovagal dizziness and fainting in response to the sight of blood or injury; Anderson, Taylor, & McLean, 1996). The patient was initially treated with standard CBT for panic disorder (Taylor, 2000). Two years later, he relapsed when exposed unexpectedly to blood-injury stimuli. This led us to hypothesize that his blood-injury reactivity played a causal role in his panic disorder. To test this possibility, we provided the patient with another course of standard CBT for panic disorder. As before, he was no longer panicking after treatment. Then we asked the patient if we could expose him to blood-injury stimuli for one month (a videotape of injections and blood extractions). The thought of being exposed to such a tape stimulated blood-injury reactions (e.g., dizziness), which were followed by a relapse of his panic disorder. The next part of the case study involved treating the patient with applied tension, which is a specific treatment for blood-injury reactivity (Öst & Sterner, 1987). This treatment reduced his panic attacks and blood-injury reactivity. When he was reexposed to the videotape, he did not have any blood-injury reactions, his panic disorder did not return, and he was free of psychopathology at his four-month follow-up. This case study involved an ABABCB design, where A = the first and second courses of CBT, B = exposure to blood-injury stimuli, and C = treatment with applied tension. This design makes it unlikely that the results are due to threats to internal validity such as history or maturation.

There are many other types of single-case experimental designs, which can be used for other types of research questions (see Barlow & Hersen, 1984; Onghena & Edgington, 2005). Studies using single-case experimental designs are useful for studying unusual cases and for conducting preliminary evaluations of new treatments. These studies are insufficient in themselves for drawing strong conclusions, but they provide some indication of whether it is useful to conduct further investigations. Early case studies of CBT for panic disorder (e.g., Clark, Salkovskis, & Chalkey, 1985) provided encouraging results, which led researchers to conduct open (uncontrolled) trials to evaluate the treatment with more patients (e.g., Sokol, Beck, Greenberg, Wright, & Berchick, 1989) and to randomized, controlled trials (which have very strong internal validity; Barlow, Gorman, Shear, & Woods, 2000) in which CBT was compared to control conditions (e.g., waiting list or placebo) and to other treatments (e.g., imipramine). Thus, even though single-case experimental designs often have far-from-perfect internal validity, they can yield valuable information and thereby can advance our understanding of psychopathology and its treatment.

Drawing inferences, whether in quasi-experiments or experiments, is a matter of ruling out rival hypotheses (e.g., hypotheses about the role of threats to internal validity) that could account for the results. Randomizing participants to experimental and control groups can overcome many of the threats to internal validity. Random selection of participants and random allocation to experimental conditions ensures, within the limits of sampling error, that the sample is representative of the target population and that the samples in the experimental groups are comparable to one another in terms of the background features of the participants, such as demographics or other variables (Cook & Campbell, 1979).

Randomization doesn't control for some threats, such as diffusion of treatments, compensatory equalization of treatments, or experimenter expectancy. These threats can be overcome by other means, such as those mentioned earlier. For quasi-experimental designs, however, there is always some degree of threat to internal validity, such as the selection threat. Cook and Campbell (1979) offer the following guidelines about how to assess the degree of threat to internal validity.

> Estimating the internal validity of a relationship is a deductive process in which the investigator has to systematically think through how each of the internal validity threats may have influenced the data. Then, the investigator has to examine the data to test which relevant threats can be ruled out. In all of this process, the researcher has to be his or her own best critic, trenchantly examining all of the threats he or she can imagine. When all of the threats can be plausibly eliminated, it is possible to make confident conclusions abut whether a relationship is probably causal. (p. 55)

External Validity

External validity has to do with the generalizability of the research findings; to what extent can the findings of an experiment or quasi-experiment be generalized *to* and *across* various populations, settings, and epochs? In the following sections, we examine, in further detail, the major types of threats to external validity, the relationship between internal and external validity, and the situations in which we should (or shouldn't) be concerned with threats to external validity. Threats to external validity are evaluated by tests of the extent to which one can generalize across various kinds of people, settings, and times and are, in essence, tests of statistical interactions (Cook & Campbell, 1979). The major threats include three types of interactions with the experimental condition that the participants are in. These are interactions with selection, setting, and history.

Interaction of Selection and Experimental Condition

Description. This concerns the question of whether the findings from the selected group of research participants can be generalized to other categories of people, such as people with other geographic or demographic features.

Examples. A study comparing patients with severe major depression with healthy controls might seek to match the participants on demographic features. Many severely depressed patients are unable to work and are therefore unemployed, receiving welfare or disability assistance. To match the patients with the controls on demographic factors, the researcher might decide to include only unemployed control participants. While this strengthens the internal validity of the study, it raises the question of whether the results can be generalized to people from other levels of occupational functioning. If the results of the research study vary across occupational levels, then there is an interaction between selection (in this case, occupational status) and experimental condition. This interaction threatens the external validity of the study. The only way to determine whether this threat exists is to determine whether the results vary with occupational status. This means that further studies might be needed to better understand the external validity of the findings.

A popular research strategy is to use analogue samples. For example, students may be selected because of their high scores on a measure of schizotypy for a study of variables thought to be relevant to schizophrenia. Analogue studies have the advantage of having strong internal validity (e.g., randomized assignment of schizotypal students to two or more experimental conditions). However, analogue studies may also have important problems with external validity. Can findings obtained from schizotypal students who, for example, report having some degree of magical thinking and perceptual aberration, be generalized to people with schizophrenia?

Studies using clinical samples also may encounter problems with external validity. Some treatment outcome studies, for example, may be highly selective in the patients that are enrolled. A study of the treatment of bulimia nervosa might only include patients if they agree to suspend any other treatment they might be receiving and remain on a stable dose of any psychotropic medication they might be receiving. These research requirements have the advantage of controlling for threats to internal validity, but they do raise questions about external validity; that is, are the patients yielding these clinical findings representative of patients typically seen in clinical practice? If the patients are not representative, then the question arises as to whether the treatment findings can be generalized to clinical practice in the "real world." These concerns with patient representativeness and the use of analogue samples were raised in the two scenarios that opened this chapter.

Even when participants belong to the target population of interest, recruitment factors might lead to threats to external validity (Cook & Campbell, 1979). A researcher, for example, who is interested in studying conversion disorder might recruit patients by placing advertisements in the local newspaper. This process of recruitment could possibly result in a sample of people with conversion disorder that is unrepresentative of people in general with this disorder. This threat to external validity can be examined by comparing patients recruited from the newspaper to patients recruited by other means (e.g., from physical referrals) to see whether the groups differ on relevant variables such as the type and severity of the conversion disorder.

Interaction of Setting or Context and Experimental Condition

Description. The question of concern is whether findings obtained in one setting or context can be generalized to other settings.

Example. Research conducted at Harvard University suggests that people who claim to have been abducted by space aliens are more susceptible, compared to control groups, to forming false memories (McNally, 2003). But do these findings apply to purported alien abductees in general, including people from other educational levels or geographic locations? Alien encounters are commonly reported in Brazil, for example (Pulos & Richman, 1990). Are these people similarly subject to false memories? To answer this question, one may need to repeat the experiment in different settings.

A related threat concerns the novelty of an intervention (Finger & Rand, 2003). If a new treatment is evaluated, typically in a university or hospital research setting, the participant may be aware that he or she is receiving a novel treatment, and the therapist may be highly optimistic or enthusiastic about the intervention. The result obtained under such conditions might not generalize to other contexts, such as settings in which the treatment is no longer regarded as novel.

Interaction of History and Experimental Condition

Description. This concerns the question of whether the findings obtained today would apply to the past or future, or whether the findings would apply to people who had otherwise different histories.

Examples. Contemporary observations of the effects of traumatic stressors and quasi-experimental analogue studies of the effects of mildly disturbing events (e.g., medical students conducting their first human dissection) suggest that exposure to traumatic events leads to symptoms of PTSD, particularly persistent reexperiencing of the event (e.g., dreams or unwanted thoughts of the event). It has been debated as to whether these are timeless responses or whether they are simply a product of contemporary Western culture. In other words, it is unclear whether the finding that stress produces reexperiencing symptoms has strong external validity. There is some suggestion that there are important limits to the external validity (generalizability) of the findings; for example, many soldiers in World War I apparently responded to the trauma of war with symptoms of conversion disorders (e.g., "hysterical" paralysis or blindness) rather than by developing reexperiencing symptoms (e.g., Lerner, 2003).

Sometimes an experiment takes place in a very special epoch, such as during the weeks following the September 11, 2001, terrorist attacks in New York and Washington, D.C. The results of a study, for example, of college student stress around the time of September 11 might not apply to other periods, past or future. One needs to rely, in part, on common sense to determine whether the results of an experiment would generalize from one time period to another.

> Even when circumstances are relatively more mundane, we still cannot logically extrapolate findings from the present to the future. Yet, while logic can never be satisfied, "commonsense" solutions for short-term historical effects lie in either replicating the experiment at different times . . . or in conducting a literature review to see if prior evidence exists which does not refute the causal relationship. (Cook & Campbell, 1979, p. 74)

Comment

Internal Versus External Validity. Internal validity often takes precedence over external validity, because one must first obtain an unambiguous finding before you can generalize the results. Accordingly, many studies in clinical psychology have high internal validity and lower (or unknown) external validity. To illustrate, in a study of memory functioning in generalized anxiety disorder (GAD), internal validity is improved if participants taking medication are excluded from the study. This is because some anxiolytic medications, such as benzodiazepines, may impair memory function. Excluding such participants improves internal validity, but it raises questions about external validity because it remains to be established that the results would apply to GAD patients who happen to be taking medication. Such patients might have clinically more severe GAD than unmedicated patients. This would be an important issue if the researcher hypothesizes that GAD arises from particular patterns of memory processing. By

excluding the more severe patients, it is not possible to determine whether the memory results can be generalized to more severe cases of the disorder.

When Does External Validity Matter? One should not automatically assume that it is important that a study has good external validity. We may not be so concerned with external validity if the focus of the investigation concerns what *can* happen, instead of what typically *does* happen (Mook, 1983). Thus external validity is less of a concern if the goal of one's research is to test predictions derived from theory or conjecture. Consider, for example, patients who report that they suddenly became aware of long-buried memories of childhood sexual abuse. The veracity of such "recovered" memories is highly controversial. Some clinicians argue that these are genuine memories that had been repressed and then retrieved. A number of researchers have argued that these are false memories, sometimes implanted by therapists using hypnosis, guided imagery, or other "memory recovery" techniques to get to the bottom of the patient's problems (for a review of this debate, see McNally, 2003). This debate raises the following question about the mechanisms of memory, which has been evaluated in several laboratory studies: Is it possible to implant a clearly false childhood memory using memory recovery techniques? Note that this is not an issue of *does* it happen but a question of *can* it happen. The answer is *yes.* Analogue research using university students has shown that it is possible to lead the participants to "recall" something that, according to their parents, never happened to them, such as being savagely mauled by a dog (e.g., Porter, Yuille, & Lehman, 1999). Although such findings have relevance to the memory recovery controversy, the primary value of this type of research is to shed light on memory processes.

To determine whether external validity is important in a given research investigation, you need to consider the conclusion that you would like to make and whether your sample and research design will enable to you reach this conclusion. The following is a sample of questions that you might ask in deciding whether the usual criteria of external validity should even be considered (Mook, 1983):

- *Regarding the Sample.* Am I trying to estimate from sample characteristics the characteristics of some population? Or am I trying to draw conclusions not about a population but about a theory that specifies what these participants ought to do? Or (as in the case of false memories) would it be important if any subject does, or can be induced to do, this or that?

- *Regarding the Setting.* Is my intention to predict what would happen in a real-life setting or target class of such settings? The answer may be *no* if the aim is to test a prediction about what ought to happen in the experimental setting. In this situation, external validity is not an issue. If the answer is *yes,* then you need to consider whether it is necessary that the setting be representative.

Evaluating and Improving External Validity. There are several ways of evaluating and improving external validity. One approach is to try to ensure that the sample is representative of the target population. The deliberate inclusion of a heterogeneous sample can be used to determine if particular variables predict the results. If you are conducting a treatment outcome study, for example, and want to know whether the results vary with socioeconomic status (SES), then you could select patients from a range of different SES levels (using stratified random sampling) and determine whether SES predicts treatment outcome. Note that this approach often requires a large sample (e.g., $n = 50$ per treatment condition), so that sufficient numbers of participants from each SES level are in each treatment condition. Another approach is to conduct multiple studies across different subgroups, settings, or times. This provides a means of determining whether the findings are replicable.

Benchmarking studies can also be used as a means of evaluating external validity. These are investigations in which research conducted in tightly controlled laboratory situations (which have high internal validity and may have low external validity) are compared with field studies (which may have good external validity but lower internal validity). A recent meta-analysis compared results from lab and field studies across a range of research domains, including clinically relevant investigations such as studies of aggression or depression (Anderson, Lindsay, & Bushman, 1999). The investigators examined the correspondence between lab- and field-based effect sizes from studies using conceptually similar dependent and independent variables. The results

of field research tended to mirror the findings from lab research, suggesting that lab studies generally have good external validity.

To illustrate benchmarking research, several studies have been conducted in which treatment-outcome findings from a university-based specialty clinic (e.g., the Center for Anxiety and Related Disorders at Boston University) are compared to findings from community mental health clinics. The university-based research tended to have high internal validity, although the use of patient inclusion and exclusion criteria raised questions about the external validity of the findings. Patients are typically excluded from CBT studies if their doses of psychotropic medication are unstable or if they have particular comorbid disorders. Studies of panic disorder, for example, often exclude patients who have comorbid paranoid, schizotypal, or borderline personality disorders. Studies conducted in community clinic settings are more liberal in their inclusion criteria and more closely approximate routine clinical treatment that patients would receive. This means that these studies have good external validity but weaker internal validity. Benchmarking studies of major depression and panic disorder indicate that the results from community clinics are similar to those obtained in university clinics and that the patients from both settings are broadly similar in their pretreatment clinical characteristics, such as the severity and duration of their disorders (e.g., Merrill, Tolbert, & Wade, 2003; Wade, Treat, & Stuart, 1998). These findings indicate that tightly controlled treatment studies from university clinics have good external validity. Such studies address the concerns of critics like Dr. Smith from Scenario 1, who claimed that treatment research findings do not generalize to patients in the "real world."

Conclusions: Perfecting Our Knowledge From Imperfect Research

Few, if any, research studies are methodologically perfect. Some consumers of the research literature tend to throw out the baby with the bathwater; that is, if a study has a minor limitation, they tend to dismiss it entirely. This was the case for the attendees of the journal club discussed in Scenario 2. But is it really true that "imperfect" studies are worthless? If this were the case, then scientific progress would not be possible—neither in psychology nor in the other sciences. But what can we legitimately conclude from imperfect investigations? Like all areas of science, no single study in clinical psychology provides the final answer to an important research question. Science is a cumulative process, whereby different studies investigate the research questions in different ways, controlling for different factors. In other words, science progresses through the development of cumulative findings from programs of research (Lakatos & Musgrave, 1970). The overall pattern of findings that emerges across studies is the most important factor in answering important research questions.

The strength of internal and external validity of a study can help researchers evaluate the relative importance of that study in an overall program of research. If a study has very weak internal validity, then it may be given little or no consideration in evaluating what the corpus of research suggests about an important research question. A study might have several strengths but might have some noteworthy weaknesses. A weakness of an analogue study of schizophrenia, for example, has the shortcoming of not using actual participants with the disorder. This is not a legitimate reason for dismissing the study altogether. The limitation simply raises another question to be answered in another study: If people who have features similar to schizophrenia (analogues) produce particular patterns of findings, then do people with full-blown schizophrenia show the same pattern of results? The analogue study may have high internal validity and lower external validity, whereas the field study (using actual patients with schizophrenia) would probably have lower internal validity (because it is difficult to control for all confounding factors when using clinical samples) but higher external validity. Together, the two types of studies complement one another.

Internal and external validity are important issues in evaluating the merits of a study, but they are not the only considerations. Other important issues include the way the data are analyzed, the reliability and validity of the measures or manipulations used, and the statistical power of the design. Those issues are discussed elsewhere in this volume.

REFERENCES

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science, 8,* 3–9.

Anderson, K. W., Taylor, S., & McLean, P. (1996). Panic disorder associated with blood-injury reactivity: The necessity of establishing functional relationships among maladaptive behaviors. *Behavior Therapy, 27,* 463–472.

Asmundson, G. J. G., Norton, G. R., & Stein, M. B. (2002). *Clinical research in mental health: A practical guide.* Thousand Oaks, CA: Sage Publications.

Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2000). Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. *Journal of the American Medical Association, 283,* 2529–2536.

Barlow, D. H., & Hersen, H. (1984). *Single case experimental designs.* New York: Pergamon.

Campbell, D. T., & Stanley, J. C. (1970). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Clark, D. M., Salkovskis, P. M., & Chalkey, A. J. (1985). Respiratory control as a treatment for panic attacks. *Journal of Behavior Therapy and Experimental Psychiatry, 16,* 23–30.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Finger, M. S., & Rand, K. L. (2003). Addressing validity concerns in clinical psychology research. In M. C. Roberts & S. S. Ilardi (Eds.), *Handbook of research methods in clinical psychology* (pp. 13–30). Malden, MA: Blackwell.

Flick, S. N. (1988). Managing attrition in clinical research. *Clinical Psychology Review, 8,* 499–515.

Furby, L. (1973). Interpreting regressions toward the mean in developmental research. *Developmental Psychology, 8,* 172–179.

Lakatos, I., & Musgrave, A. (1970). *Criticism and the growth of knowledge.* Cambridge, UK: Cambridge University Press.

Lerner, P. (2003). *Hysterical men: War, psychiatry, and the politics of trauma in Germany, 1890–1930.* New York: Cornell University Press.

Margraf, J., Ehlers, A., Roth, W. T., Clark, D. B., Sheikh, J., Agras, W. S., et al. (1991). How "blind" are double-blind studies? *Journal of Consulting and Clinical Psychology, 59,* 184–187.

McNally, R. J. (2003). *Remembering trauma.* Cambridge, MA: Harvard University Press.

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology, 71,* 404–409.

Mogg, K., & Bradley, B. P. (1999). Selective attention and anxiety: A cognitive-motivational perspective. In T. Dalgelish & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 145–170). New York: Wiley.

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

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21,* 56–68.

Öst, L.-G., & Sterner, U. (1987). Applied tension: A specific behavioral method for treatment of blood phobia. *Behaviour Research and Therapy, 25,* 25–29.

Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted, and fabricated childhood emotional events: Implications for the recovered memory debate. *Law and Human Behavior, 23,* 517–537.

Pulos, L., & Richman, G. (1990). *Miracles and other realities.* Vancouver, British Columbia: Omega.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist, 57,* 839–849.

Sokol, L., Beck, A. T., Greenberg, R. L., Wright, F. D., & Berchick, R. J. (1989). Cognitive therapy for panic disorder: A nonpharmacological alternative. *Journal of Nervous and Mental Disease, 177,* 711–716.

Taylor, S. (1994). The overprediction of fear: Is it a form of regression toward the mean? *Behaviour Research and Therapy, 32,* 753–757.

Taylor, S. (2000). *Understanding and treating panic disorder.* New York: Wiley.

Taylor, S., Thordarson, D. S., Maxfield, L, Fedoroff, I. C., Lovell, K., & Ogrodniczuk, J. (2003). Comparative efficacy, speed, and adverse effects of three treatments for PTSD: Exposure therapy, EMDR, and relaxation training. *Journal of Consulting and Clinical Psychology, 71,* 330–338.

Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking study. *Journal of Consulting and Clinical Psychology, 66,* 231–239.