Outcomes for Children and Youth with Emotional and Behavioral Disorders and Their Families

Programs and Evaluation Best Practices

SECOND EDITION

Edited by

Michael H. Epstein
Krista Kutash
Albert J. Duchnowski

pro-ed
An International Publisher

8700 Shoal Creek Boulevard
Austin, Texas 78757-6897
800/897-3202 Fax 800/397-7633
www.proedinc.com

pro-ed
An International Publisher

8700 Shoal Creek Boulevard
Austin, Texas 78757-6897
800/897-3202 Fax 800/397-7633
www.proedinc.com

Library of Congress Cataloging-in-Publication Data

Outcomes for children and youth with emotional and behavioral disorders and their families : programs and evaluations, best practices / edited by Michael H. Epstein, Krista Kutash, Albert Duchnowski.—2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-89079-989-X (sc. : alk. paper)
1. Behavior disorders in children. 2. Adolescent psychopathology. I. Epstein, Michael H. II. Kutash, Krista. III. Duchnowski, Albert J.

RJ506.B44098 2004
618.92'89—dc22

2004044232

Art Director: Jason Crosier
Designer: Nancy McKinney-Point
This book is designed in Gill Sans and Minion.

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10 06 05 04 03 02

Research Designs for Children's Mental Health Services Research

CHAPTER 4

Stephanie Reich and Leonard Bickman

Research in children's mental health services is focused on determining what treatments and methods for delivering services work best for children with mental health needs. To answer these questions, researchers often employ experiments and quasi-experiments in an attempt to identify a causal relationship between variables (e.g., Does this treatment result in better outcomes for children?). There are many different ways to study a causal relationship, and each method has potential problems associated with it. This chapter describes the criteria needed to determine causality and the key features of experimentation and quasi-experimental designs, and then enumerates the many issues that threaten the validity of a study and potential interpretations of the study results.

DETERMINING CAUSALITY

To determine that a causal relationship exists, several criteria must be met. First, the temporal order must be that the cause precedes the effect. Second, there should be temporal contiguity between the cause and effect; that is, the cause and the effect appear relatively close in time. Third, there should be a level of common variation, known as covariation, between the cause and the effect. When the cause is present, the effect is present, and when the cause is absent, the effect is absent. Fourth, there should be congruity between both. Typically, a small cause should result in a small effect, whereas a large cause should result in a large effect. When there is a mismatch (e.g., small cause produces big effect), more evidence is usually needed to explain this discrepancy (Cordray, 2000). Finally, plausible rival explanations of the observed effect should be improbable. Ruling out other explanations supports the proposed causal relationship. This is accomplished best by addressing the threats to validity described later in the chapter.

COUNTERFACTUAL CONDITION

The ideal way to test a causal relationship is to study a person or situation with an intervention and without it. This enables a research investigation to determine the effect of the cause by comparing what happened with the

intervention to what would have happened without it. The "what would have happened" condition is called the *counterfactual condition* (Corrin & Cook, 1998; Holland, 1989) and is based on the assumption that everything else is the same except for the presence of the intervention (Shadish, Cook, & Campbell, 2002). In the real world, testing the counterfactual condition is not possible. Often the intervention has a lasting effect that would influence a later studied no-treatment condition. Participants who experience an educational intervention could not easily "unlearn" the program in order to be measured again in a no-treatment condition. Even if the intervention effects could be removed, there are many other characteristics that would be altered. The time period would be different and the participants would be more experienced with the study. This is why using each person as his or her own control is fraught with problems. Typically, the closest a researcher can come to the counterfactual condition is to use a comparison group.

A comparison group is a group of people who do not receive the "cause" (i.e., treatment or intervention) and who are compared to the group that does (treatment group). The difference between these groups, if everything else is the same, is the effect of the intervention. The more similar the comparison group is to the treatment group, the greater the confidence in the causal relationship or that the effect was produced by the cause.

EXPERIMENTS AND QUASI-EXPERIMENTS

Two ways researchers can investigate a causal relationship is through a randomized experiment or a quasi-experiment. In a randomized experiment, the researcher randomly assigns participants into groups, levels, or conditions such that each participant has an equal chance (i.e., independent, nonzero probability) of being assigned to any of the experimental groups, levels, or conditions (Keppel, 1991). In quasi-experimental designs, participants are not randomly assigned to conditions.

Random assignment reduces the possibility of systematic bias from characteristics of the participants appearing more often in one group than the other (Hill, Rubin, & Thomas, 2000). Participants may vary in many ways and some of these characteristics may affect the behavior (effect) being studied. Without random assignment, these differences may be associated with how a person is placed in a group. When random assignment is carried out successfully, there is greater confidence that the treatment and control groups were equivalent on both measured and unmeasured variables before the treatment was initiated.

Imagine we believe that caregivers of children with severe emotional disturbance (SED) feel additional stress by having chaotic schedules in which they have to juggle their own agenda with their children's school schedule and mental health appointments. In response, we create a program that is designed to help people better manage their time. We decide to conduct

experiment in which half of the caregivers receive the program and the other half do not receive assistance with scheduling. We ask all 40 parents to come to the study facility at one specific time and we assign them to groups on a first-come basis. The first 20 parents to arrive are placed in the treatment group and the next 20 are assigned to the no-treatment group. At the end of the program we find that our treatment group members, who arrived earlier, are much more organized with their time than the other group. Is the difference observed at the end of the study because of the program or some other factor? Perhaps the caregivers who are punctual are already better at scheduling their time than the late-arriving ones are. Another explanation accounting for the difference could be that more of the people who arrived on time have cars, which provide more flexibility in scheduling, whereas people who arrived later rely on the bus and have less freedom in scheduling. If we had randomly assigned caregivers to groups, it would be more likely that the early- and late-arriving people would be equally distributed among the groups. Instead, we may have introduced systematic bias into our study.

Randomization works best when applied to large groups of participants. For example, if you flip a coin once, you have a ½ chance of getting a head. If you flip it five times, you have a $(½)(½)(½)(½)(½) = 1/32$ chance of getting all heads. The more people in the study, the less likely similar characteristics will disproportionately appear in one group more than the other. A researcher must be sure that the random assignment process is truly random and not simply haphazard. In an evaluation of a treatment, it is often better if the researcher, rather than the people delivering the services, conducts the random assignment. This helps ensure that the randomization process is not corrupted, for example, by assigning more "needy" cases to the treatment group.

A key factor in maintaining the credibility of a study design is that others consider it to be as unbiased as possible. When feasible, it is best to collect baseline (pretest) data prior to randomly assigning participants to groups. This helps ensure that the groups are equivalent because early dropouts (persons who do not complete the baseline data collection) will have dropped out before random assignment and thus will not contribute to differences between the groups. The data collected can also provide information on persons who drop out early. Moreover, if there is a key variable that the researcher wants to have randomly distributed, baseline data will allow for pairing participants on that variable and then randomly assign each member of a pair. Random assignment does not guarantee equivalent groups; it just increases the probability that influential variables will be equally distributed between groups rather than disproportionately clustering in one. In the next section, we describe a randomized study of children's mental health services to illustrate these points.

All research, including the examples cited in this chapter, requires ethical considerations in its design and implementation. Research procedures should be designed to protect the privacy and safety of participants. Each of the examples of children's mental health services research in this chapter was reviewed and approved by a university institutional review board (IRB), which has the express purpose of protecting the rights of children and their

families. For more information on ethical conduct for research with human participants, please see Appendix 4.A.

The Stark County Study: Example of a Randomized Experiment

The Stark County study was an evaluation of a comprehensive system-of-care model using a randomized design (Bickman, Summerfelt, & Noser, 1997). The term *system of care* is a concept of integrated comprehensive care for children and was developed in response to problems concerning the availability and delivery of mental health services for children; this framework is discussed elsewhere in this volume. To recruit participants for the study, names and contact information from families seeking services were obtained from either the Department of Human Services or the local community mental health center. The center intake worker followed the clinic's usual guidelines to determine eligibility for services. Children were excluded from the evaluation if they were too young (less than 5 years old), did not have an SED, had too low an IQ (less than 85), or if the admitting center personnel considered the child to be in need of emergency services. The clinic intake staff, based on their usual procedures, made the determination of SED and also asked for permission for the Vanderbilt University evaluators to contact the parent. When contacted by the evaluation staff, 85% of the parents agreed to participate in the study and be interviewed.

Only after the interview data were collected did a computer program randomly assign the family to either the system of care (treatment group) or usual care from the community (comparison group). In this latter group, caregivers were told that they had to arrange services for their children on their own and were given a list of community providers.

The evaluation staff carried out the random assignment process after baseline data had been collected, reducing the potential threats to validity caused by differential attrition (e.g., families drop out at different rates from the treatment and comparison groups). The primary advantage of random assignment in this study was the increased probability of initially equivalent groups. However there were also potential problems associated with such a design; for example, critics of the evaluation claimed that the comparison group children were deprived of better services. The sole purpose of the study was to determine if a system-of-care model was more effective than treatment as usual. This model of care had not been tested and whether it was better was unknown. Therefore, it could have been argued that it was more unethical to continue to provide untested services to children and families than it was to withhold experimental services of unknown effectiveness.

Another potential problem with this experimental design was that although random assignment was used, it was not employed for all possible participants. Certain children were excluded from the study because they were in need of immediate care. Random assignment for such cases was not

possible because collecting pretest data before assignment would have delayed treatment. As will be discussed later, the exclusion of this group limits the generalizability of the findings to nonemergency cases but does not affect the researcher's ability to make a conclusion about causality. Although randomization is often preferable, there are times when it is inconvenient, impractical, or unethical (Lipsey & Cordray, 2000).

At times, randomizing participants is burdensome and expensive. In such instances, relying on naturally occurring groups is more practical. School-based programs often utilize comparisons of students between schools rather than trying to randomly assign children to schools. At other times, randomization is simply not possible. If a program has full coverage, in which every member of the population receives the treatment, then it is not possible to randomly assign people to a treatment or no-treatment comparison group. This is often the case when entitlement policies are studied or if the study deals with a natural disaster (Hendrick, Bickman, & Rog, 1993). For example, if federal law entitles a certain class of people to services, it would be illegal to deny services to anyone in that group. Randomization is also impossible if a condition occurs in only a subset of the population. For example, a researcher cannot assign people to have major depression or to be female rather than male. It is difficult to imagine assigning someone to the conditions of poverty, incarceration, or homelessness. Finally, at times randomization may be possible but unethical. A researcher comparing two different effective types of services in order to determine which is more effective could not add a no-treatment control condition because an effective treatment exists. However, if there is no effective treatment, random assignment to a no-treatment control condition is acceptable. In practice, there is no ethical way to control whether people obtain services outside of the study. Thus, the comparison group is often labeled a treatment as usual group (TAU) rather than a no-treatment control group. More information about the use of randomized designs can be found in Boruch (1997).

When random assignment is not possible due to expense, practicality, or ethics, a researcher must rely on quasi-experimental designs. Although there are several types of quasi-experimental research designs, the three most relevant to mental health services are the (a) nonequivalent comparison group design, (b) the interrupted time series design, and (c) the regression discontinuity design.

Nonequivalent Comparison Group Design

When randomization is not possible, a researcher can use a nonequivalent comparison group design in which groups are identified by some naturally occurring criteria. This design is the most common quasi-experimental design used in the behavioral sciences (Rosenthal & Rosnow, 1991). Nonequivalent group designs can compare the treatment group to a deliberately chosen group, normed data from another group/sample, or secondary data collected

from other studies (Shadish et al., 2002). Deliberately selecting a comparison group that is most similar to the treatment group is often the best method for making causal inferences.

Technically, the nonequivalent group design could include measurement as a pretest (before the intervention) and posttest (after the intervention) or only as a posttest (Reichardt & Mark, 1998). Because the groups were determined on some criteria other than randomization, using only posttest data renders causal explanations fraught with problems. Without measuring baseline status, it is difficult to conclude that differences between the groups were due to the intervention rather than to another factor that existed before the intervention.

The Fort Bragg Evaluation: A Nonequivalent Comparison Group Quasi-Experiment

The largest nonequivalent comparison group study evaluating child and adolescent mental health services was the Fort Bragg Demonstration Project and Evaluation (Bickman, 1996a, 1996b; Bickman et al., 1995). The Fort Bragg study was designed as an independent evaluation of the Fort Bragg Child and Adolescent Mental Health Demonstration to test the efficacy of providing a full continuum of community-based services. The goal was to determine if this continuum of services resulted in improved treatment outcomes while decreasing the cost of care when compared to treatment as usual.

The evaluation results revealed that the demonstration successfully implemented a continuum of care and that it dramatically increased access to mental health services. However, children at both the demonstration and comparison sites improved on measures of mental health outcomes. Children in the demonstration showed no greater improvement than did children at the comparison site, and the costs of services at the demonstration site were higher than at the comparison site. Instead of confirming strongly held beliefs, the evaluation reported that the continuum was less cost-effective than the fragmented treatment found in the comparison site.

The initial intent was to use a randomized design for the Fort Bragg evaluation. However this was not possible because the system-of-care intervention was a full-coverage program affecting all children in Fort Bragg, North Carolina. Every family in the area had access to the demonstration; therefore, none of the families could be assigned to a comparison group. Because the project was unable to use random assignment of participants to different models of care, families and children from other military bases served as comparison participants to facilitate examination of the effectiveness of the demonstration. Army officials designated two comparison sites that were approximately the size of Fort Bragg: Fort Campbell, Kentucky, and Fort Stewart, Georgia. Both bases provided children with traditional mental health services.

One of the major challenges with quasi-experimental designs is ensuring that the treatment and comparison groups are equivalent at the start of the study. Although modern statistical techniques make potential lack of equiva-

lency less of a problem than in the past, differences at the start of the study clearly complicate interpretation of causal relationships. The study therefore needed to assess whether the families and the service settings were similar before the introduction of the demonstration.

The Fort Bragg evaluation was fortunate to have military posts as the sites of both the treatment and comparison conditions, because separate posts tend to be quite similar. In contrast to families in different cities or different parts of a city, military families and posts are very comparable regardless of where they are located.

However, these similarities did not guarantee that the families and children from each site would be the same on important mental health variables. Therefore, the evaluation staff compared the children and families on 103 mental health variables at baseline. This large number of variables was selected because the investigators wanted the comparisons to be exhaustive and include all subscales of the instruments. The sites differed statistically on 14 variables, with the comparison site children appearing more impaired on 9 variables and the demonstration children seeming more impaired on 5 other variables. These differences in groups were very small, and the researchers concluded that it was unlikely that they would account for differential outcomes in the posttest. Because this was not a randomized experiment, the researchers could not be sure that the groups did not differ on some important variable that they did not measure. It was believed, however, that this was unlikely given the number and importance of the variables tested.

Thus, although a randomized experiment could not be conducted, the quasi-experiment had the advantage of being able to include all families, even families in crisis, in contrast to the randomized Stark County study, discussed earlier, which eliminated this group of families. Another advantage of this design was the lack of delay in treatment because children were recruited as soon as they entered services. In a randomized experiment, there may be a delay in the receipt of services if the researcher wants to collect pretest data before the child is assigned to either the treatment or the comparison group. As noted earlier, children who are in need of immediate services could not participate in a randomized study (such as Stark County) but could be part of the quasi-experiment. In such cases where there is a choice of designs, a researcher will have to compare the benefits of randomized versus nonequivalent comparison group designs. These trade-offs address issues of internal and external validity of designs and will be discussed in greater depth later in the chapter.

Interrupted Time Series Design

When there are many data collection periods before and after an intervention, the quasi-experimental design is known as an interrupted time series design, which may or may not include a nonequivalent comparison group. A time series design is the repeated measure of a variable over time. An interrupted time series design is the repeated measurement of a variable before and after the introduction of a new variable, such as a treatment or intervention. The

more often the variable is measured over time, the more sensitive the study will be to changes caused by the introduction of the treatment. By determining the typical trend of the variable before and after the intervention, researchers can identify the degree of impact, or effect, of the intervention. Typically, estimating the effect can be determined by calculating the slope or rate of change before and after the intervention. Data collection may look like this:

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad X \quad O_5 \quad O_6 \quad O_7 \quad O_8$$

Os indicate the measurement periods and *X* indicates the introduction of the intervention.

The change in the data may look like a large level shift, a slow increase or decrease, a temporary change, a delayed change, or an abrupt change that decays (Reichardt & Mark, 1998). Figure 4.1 demonstrates some of these changes. Interrupted time series designs allow the treatment group to serve as

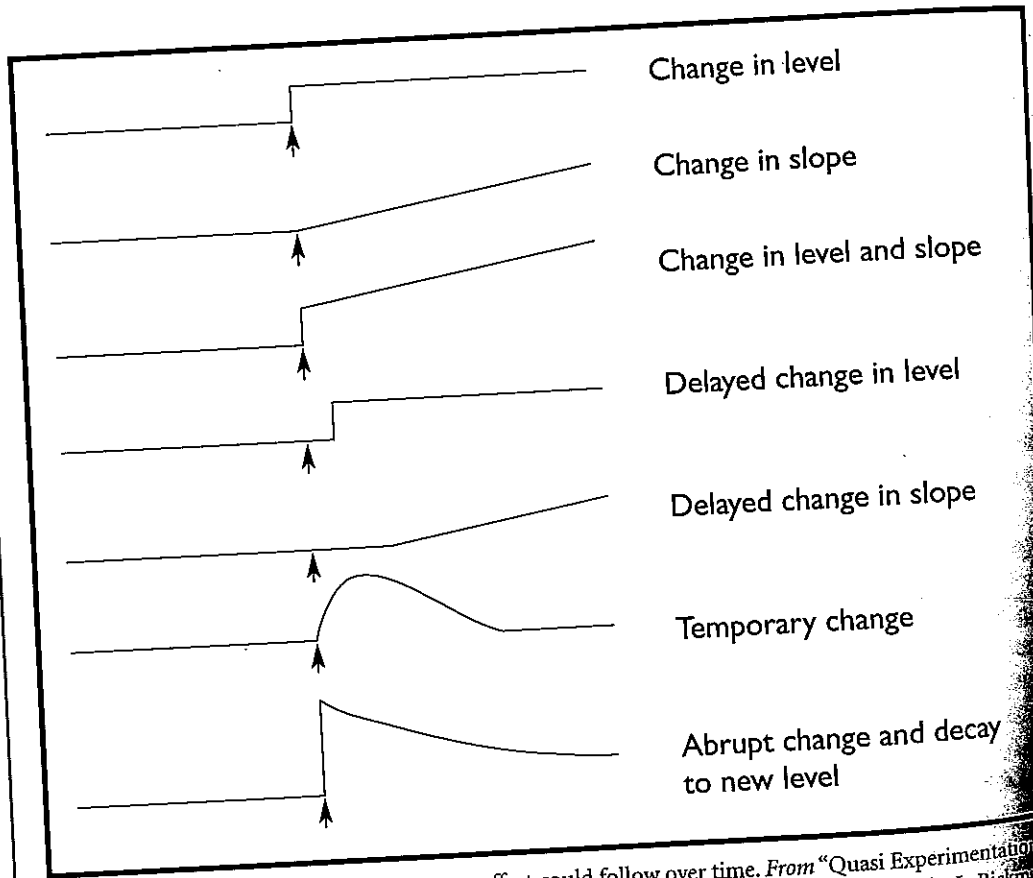

FIGURE 4.1. Possible patterns that a treatment effect could follow over time. *From* "Quasi Experimentation" by C. Reichardt and M. Mark, in *Handbook of Applied Social Research Methods* (pp. 193–227) by L. Bickman and D. Rog (Eds.), 1998, Thousand Oaks, CA: Sage. Copyright 1998 by Sage. Reprinted with permission. This figure was originally adapted from "Estimating Effects of Community Prevention Trials: Alternative Design and Methods," by C. Reichardt, in *Community Prevention Trials for Alcohol Problems: Methodological Issues* (pp. 137–158) by H. D. Holder and J. M. Howard (Eds.), 1992, Westport, CT: Praeger. Copyright 1992 by Greenwood Publishing Group, Inc. Adapted with permission.

its own comparison group. Multiple measurement points before and after the intervention allow researchers to compare how the measured variables were changing over time without the intervention and how they changed after the intervention. Interrupted time series designs can use a nonequivalent comparison group as well. In such situations, both the treatment group and the comparison groups are measured at multiple times before and after the intervention is introduced to the treatment group. Such a design would look like the following:

Treatment group:	O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8
Comparison group:	O_1	O_2	O_3	O_4		O_5	O_6	O_7	O_8

Wraparound Services: An Interrupted Time Series Example

In 1996 Congress mandated that the Department of Defense (DoD) develop, implement, and evaluate a demonstration project that utilized a "wraparound" service system for child and adolescent military dependents. Congress defined "wraparound" as a community-based program developed with a focus on individual needs to support normalized and inclusive options for child and adolescent mental health patients and their families. The evaluation addressed child and family outcomes and costs. A randomized experiment was not acceptable to DoD, so a nonequivalent comparison group design was used (Bickman, Smith, Lambert, & Andrade, 2003). The treatment group contained the children and families who entered the wraparound program. However, a comparison group was not easily established. The evaluators chose to recruit families who had been referred to the demonstration but did not participate in it. The logic was that the groups should start out equivalent because the children were referred by the same sources. However, it was possible that the demonstration's eligibility criteria and the self-selection of the families into the demonstration could produce two very different groups of children. Therefore, a comparison of the baseline data of those who agreed to participate in the evaluation and those who did not was conducted; this showed only 5 statistically significant differences out of 90 variables (which would be expected by chance alone).

Although the clinical outcomes for this study were evaluated in a nonequivalent group design, the cost analysis used a form of the interrupted time series design with a nonequivalent comparison group. This design is very useful for studying variables like service use or cost, which are usually collected frequently.

For service utilization and the cost analysis of the wraparound program, data from the Health Care Services Record (HCSR) were used. HCSR data report the volume and type of services children received and the dollar amount providers billed for those services. Summaries of these data were supplied for the services children received over the 3 years preceding the start of the wraparound demonstration and ended 1 month before any child received the last

wraparound service. A summary was done per child for each month of care. On average, the HCSR data contained 48 months of service utilization and mental health expenditures per child.

When analyzing monthly HCSR information, we used a nonlinear longitudinal hierarchical model known as a piecewise linear model (PWLM). The main advantage of using a PWLM is that it allows measurement of changes in expenditures between groups across different time segments (Lambert, Wahler, Andrade, & Bickman, 2001; Snijders & Bosker, 1999). The results indicated that both groups had equivalent cost histories for the 30 months before the wraparound demonstration. In the 6 months preceding the demonstration, both groups experienced an equivalent dramatic rise in costs. However, in the 6 months following the start of the demonstration, costs dropped more dramatically for the comparison group than for the wraparound demonstration group, resulting in significantly lower costs for the comparison group. As in the two studies discussed above, child and family outcomes were equivalent but costs were higher in the experimental condition.

Regression Discontinuity Design

One of the risks of not randomly assigning participants to groups is that unmeasured characteristics may bias the effects. However, if we attempt to understand the criteria for how people are assigned to groups, we can attempt to control for their influence. Regression discontinuity designs determine a priori the criteria for how people are assigned to groups. Therefore, the selection process, as with random assignment, is known perfectly, at least in theory (Cook & Campbell, 1979).

In a regression discontinuity design, a researcher creates or identifies an assignment variable, which can be any measure taken at baseline. Once a measure is identified, the researcher determines a critical cutoff point. Participants with scores above this point are assigned to one group, and those below are assigned to the other. The cutoff point should be precise so that variations around this point are not mistaken for effects (Cook & Campbell, 1979). In a regression discontinuity design, the method for assignment can be perfectly measured and implemented (Shadish et al., 2002). Figure 4.2 illustrates the implementation of a regression discontinuity design. This design would be very useful if a service used objective measures and consistent cutoff points to assign different levels of care. However, this type of decision making is rare in mental health services, and this design is not frequently used.

THREATS TO VALIDITY

All experimental designs, including quasi-experiments and randomized experiments, are susceptible to external variables that can threaten the conclusions of a study. Inferences (e.g., conclusions and generalizations) about the findings are vulnerable to the threats in four categories of validity: (a) statisti-

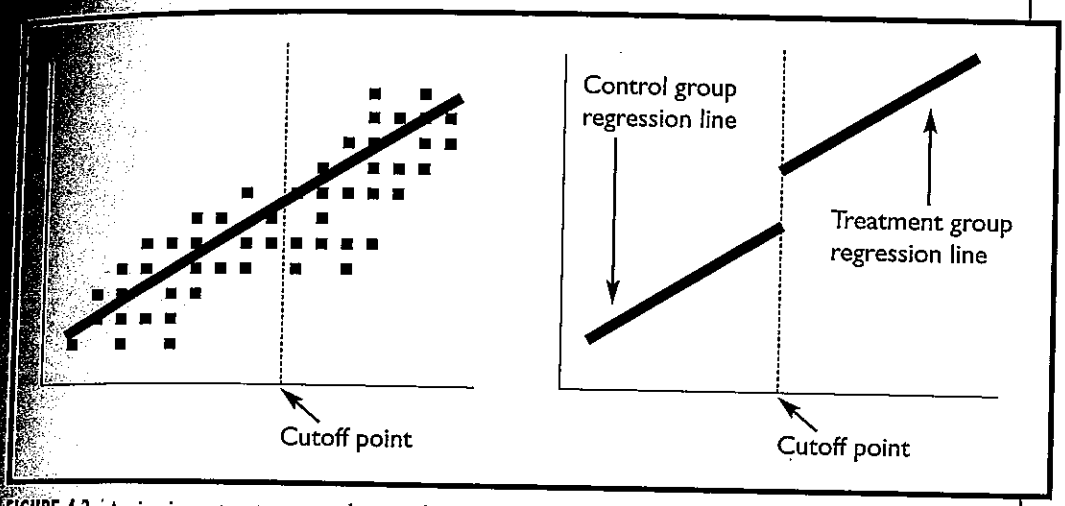

FIGURE 4.2. Assigning a treatment and control group in a regression discontinuity design.

cal conclusion validity, (b) internal validity, (c) construct validity, and (d) external validity (see Table 4.1). Internal validity deals directly with the causal relationship: Did A cause B? Statistical conclusion validity determines if there is quantitative evidence of a covariation between A and B. Construct validity deals with inferences made from findings: Are A and B conceptually what the researcher believes them to be, or has the relationship been mislabeled? External validity describes what applications of the causal relationship can be made. Specifically, can finding that A causes B be applied to other people, places, or times than this study? Understanding how these threats operate is critical to designing good studies. Because the goal of the good designer is to eliminate as many of these threats as possible, they are described in some detail in this chapter.

These four threats to validity apply to randomized experiments and noncausal relationships (e.g., correlations), but quasi-experimental designs are often more vulnerable to their effects. Therefore, this section will focus on how threats to statistical conclusion, internal, construct, and external validity can threaten the interpretations of causal relationships in quasi-experimental designs.

Statistical Conclusion Validity

If a study has statistical conclusion validity, then "conclusions about covariation are made on the basis of statistical evidence" (Cook & Campbell, 1979, p. 37). In quasi-experimentation, these conclusions are drawn about the cause and effects on the basis of proper statistical methods applied appropriately to reliable information (Wortman, 1994). Basically, statistical procedures can demonstrate that the cause is related to a change in the effect. To assess the statistical conclusion validity of a study, one must determine if (a) the study design is sensitive enough to detect covariation between two variables (i.e.,

TABLE 4.1
Threats to the Four Types of Validity

Statistical conclusion validity
- Low statistical power
- Fishing and error rate
- Unreliability of measures
- Unreliability of treatment implementation
- Random irrelevancies in the setting
- Random heterogeneity of respondents
- Violation of statistical tests

Internal validity
- History
- Maturation
- Testing
- Instrumentation
- Statistical regression
- Selection
- Attrition (mortality)
- Demoralization
- Contamination
- Compensation
- Interactions with selection
- Ambiguity about the direction of the causal inference

Construct validity
- Inadequate preoperational explication of the constructs
- Mono-operational bias
- Mono-method bias
- Hypothesis guessing
- Evaluation apprehension
- Experimenter expectancies
- Confounding of constructs
- Interaction of different treatments
- Interaction of testing and treatment
- Restricted generalization across related constructs

External validity
- Interaction of selection and treatment
- Interaction of setting and treatment
- Interaction of history and treatment

cause and effect), and (b) if so, what the evidence of the covariation is, and (c) given the evidence, how strong the covariation is (Rosenthal & Rosnow, 1991).

There are seven major threats to statistical conclusion validity that make drawing conclusions about covariation difficult. These are (a) low statistical power, (b) fishing and error rate, (c) unreliability of measures, (d) unreliability of treatment implementation, (e) random irrelevancies in the setting, (f) random heterogeneity of respondents, and (g) violation of statistical tests (Cook, Campbell, & Peracchio, 1990).

Statistical Power

Experimentation is based on supporting or rejecting the null hypothesis. The null hypothesis is the hypothesis that would be true if the experimental one is false. The null hypothesis is assumed to be true in generating the sampling distribution used in the study (Hays, 1994). For example, if a study hypothesizes that A causes B, the null hypothesis would state that there is no causal relationship between A and B. Thus, if the study fails to reject the hypothesis, then it cannot conclude that there was an effect (B) produced by a cause (A). On the other hand, if the study is able to reject the null hypothesis, it has demonstrated that there appears to be a relationship between the two variables.

In testing causal relationships, there are four possible findings: (a) that a relationship between the cause and effect exists and the study statistically identifies it, (b) that no such relationship exists and the study incorrectly concludes that there is one (false positive), (c) that no relationship exists and the study indicates no relationship, and (d) there is a relationship and the study falsely concludes that there is none (false negative). A Type I error occurs when covariation is found to be statistically significant but in actuality there is no relationship. The Greek letter α (alpha) represents the probability of such a false positive. When there is covariation and the researcher does not identify it (i.e., false negative), this is known as a Type II error. The Greek letter β (beta) represents the probability of making this error. *Statistical power* is the probability of identifying the covariation between the cause and effect when it is does, indeed, exist. This probability is $1 - \beta$. Several factors affect the statistical power of a study. These are the sample size (number of people in each condition of the study), alpha level (the probability of a false positive), the type of statistical test applied, and the effect size (the difference between the means of the treatment and comparison groups divided by their standard deviation; Lipsey, 1998).

Low statistical power is a threat to statistical conclusion validity because it increases the likelihood of falsely concluding that there is no relationship between the cause and the effect. Low statistical power is similar to saying a study is not very sensitive. To guard against this, the researcher should increase sample sizes, raise the alpha level, use more powerful statistics, or assess larger effect sizes. It is absolutely critical that statistical power be calculated during the design of the study. If the power is not sufficient to detect the expected effect size, it may not be wise to conduct the study.

Fishing and Error Rate

The alpha level is the probability of making a false positive. In the behavioral sciences this is typically set at $p < .05$. This means that out of every 100 studies conducted, 5 will be found significant by chance alone. Fishing occurs when multiple analyses of the data are conducted without adjusting the significance level. The more statistical comparisons that are conducted in a single study, the greater likelihood chance will bias the results. This threat can be decreased by reducing the number of analyses conducted, making statistical adjustments

(corrections that lower the significance level, such as the Bonferroni procedure) when performing multiple analyses, and using multivariate techniques when testing relationships among many variables (Cook et al., 1990).

Reliability of Measures and Restriction of Range

The reliability of measures poses a particular threat to statistical conclusion validity. Reliable measures consistently measure a construct (Shadish et al., 2002). If a measure is unreliable, unintended variation is added to the research design. Unreliability may lead to nonsignificant findings when there was actually an effect produced. Therefore, unreliable measures threaten validity in either direction (i.e., false negative or false positive; Rogosa, 1980). Additionally, measures that restrict the range of data will not be sensitive to changes in participants. This is likely to occur when a large proportion of participants have extremely low or extremely high scores. They either cluster near the lowest possible score (floor effect) or at the highest possible score (ceiling effect). In either case, it is difficult to measure change because scores are already at the range limits and can be neither lower nor higher (Shadish et al., 2002).

Reliability of Treatment Implementation

To test whether the cause (e.g., treatment) produces an effect, it is important that the cause be implemented consistently. This reliability in implementation allows for greater causal inferences. For example, conclusions about the effects of a system-of-care model for children's mental health would be difficult to draw if some sites did not apply the model or only did so sporadically. Low treatment reliability or fidelity could also occur if one site used the system-of-care model for some families and not others within the same study.

Extraneous Variance in Setting

Variations in the setting in which the treatment occurs may have an influence on the effect. These influences can be any type of environmental factor (e.g., time of day, temperature, noise) that varies at different time points or between the comparison and intervention groups. This added variance makes it more difficult to identify which effects were produced by the intervention rather than the setting. Studies conducted within laboratory settings use a much smaller sample of participants to detect smaller effects than "real-world" field-based studies because the researcher or investigator can control and reduce the effects of setting variation. The ability to control the setting is almost absent in studies of treatment or services that take place in the real world.

Random Heterogeneity of Respondents (Units)

If the participants in a study are very different from one another, they may produce findings with a great deal of variation that is not due to the treatment. This extra variance may obscure covariation between the cause and

fect, either by exacerbating or minimizing it. For example, laboratory studies of psychotherapy usually reduce variability by severely limiting the participants to a single diagnosis, a luxury that is not available in a community-based mental health center or clinic.

Violation of Statistical Test

Statistical conclusion validity is based on the statistical evidence of covariation between two variables. This evidence is obtained through statistical analysis. Each statistical test has certain assumptions that must be met for it to be valid. These assumptions should be known before the researcher applies them to data. False conclusions may be drawn if an inappropriate test is used to assess covariation. For example, analysis of variance (ANOVA) is intended for normally distributed data. If the data are highly skewed or bimodal, transformation may need to be conducted first or perhaps a different analytic procedure applied.

Statistical conclusion validity, although very important for assessing covariation and drawing causal inferences in quasi-experimental and experimental designs, has been written about the least in the research methodology literature. However, there is growing concern over this lack of awareness (Lipsey, 2000). Specifically, a lack of statistical power is an especially large threat to statistical conclusion validity. In 1962, Cohen found that social research had enough power to detect expected effects only half the time (Cohen, 1962). Twenty-seven years later, Sedlmeier and Gigerenzer (1989) found little improvement in the social sciences. Overall, most social science is underpowered, meaning that it is not sensitive enough to detect important smaller effect sizes. Clearly, there is no point in conducting well thought out and designed studies if they are not powerful enough to identify covariation when it is present. To avoid this, all planning for experimentation (including quasi-experimentation) should include a power analysis. Otherwise, treatments may be labeled as ineffective when, in actuality, they may be of benefit. Consumers of research should be aware of issues of power when reading studies that conclude no effect. Readers should be cognizant of sample sizes, alpha levels, the types of statistical tests conducted, and the effect sizes anticipated by the researchers. (For more information on statistical power, see Cohen & Cohen, 1983; Lipsey, 1990.)

Internal Validity

Internal validity refers to the truthfulness "with which statements can be made about whether there is a causal relationship from one variable to another" (Cook & Campbell, 1979, p. 38). Many factors can threaten the internal validity of a design. These include history, maturation, testing, instrumentation, statistical regression, selection, attrition (mortality), demoralization, contamination, compensation, interactions with selection, and ambiguity about the direction of the causal inference. Each of these will be briefly discussed.

History

Events other than the treatment that have an influence on participants' behavior can threaten internal validity (Shaughnessy & Zechmeister, 1994). These events typically occur after the pretest and before the posttest so that changes or lack of changes could be due to the intervention or some other event. A historical event may exaggerate or cloud findings (McCleary, 2000). Imagine the implementation of a school-based program to identify signs of depression and suicidal risk in teens. At the same time the program is being implemented, the city begins to air public service announcements detailing the warning signs of suicide and providing information on services. The posttest of the school-based program shows that there is an increase in staff and students' identification of depressive and suicidal behaviors. With these findings, it is difficult to conclude whether the program worked, the changes were due to the commercials appearing nightly, or the changes were due to a combination of the program and the announcements. Historical events thus make causal inferences difficult. Historical threats can apply to randomized designs if one group is exposed to some influencing event that the other group is not.

Maturation

Studies that occur over time are susceptible to the influence of historical changes as well as changes within participants that were not intended by the treatment. These changes could include growing older, stronger, wiser, more fatigued, or bored. If the changes affect the intended outcome, inferences about the effects of the treatment are difficult. This threat to internal validity is a common concern when researchers are working with young children. Because children grow and change quickly, it is often more difficult to separate changes due to natural maturation from effects of the intervention. For example, in a program to treat hyperactive and inattentive behaviors in young children, significant findings may be attributable to either the program or to the fact that the children have matured and display more age-appropriate self-regulation. This is a serious concern for research in the field of child and adolescent mental health services.

Testing

Testing participants multiple times also may affect the answers they give. This is called *reactivity.* Reactivity to testing can occur in several ways. Sometimes participants want to be consistent with previous answers, thus minimizing changes over time. The test may pique participants' curiosity about something on the pretest and provoke them to look up the answers or at least think more about the topic. Both thinking more and obtaining information may improve participants' performance on the subsequent testing. In such cases their improvements are not due to the intervention but to their own studying. These reactive testing effects are sometimes referred to as *practice effects.*

There are several ways to assess the impact of testing effects. Item response theory (IRT) enables researchers to calibrate participants' performance through the use of multiple measures (Lord, 1980). Additionally, research

signs can detect the influence of testing effects by providing comparisons between groups with and without pretest measures. The Solomon four-group design is the best example of this (Hendrick et al., 1993; Solomon, 1949). This design utilizes four groups: two intervention groups and two comparison groups. One of the intervention groups and one of the comparison groups are given a pretest, and all four groups are given a posttest. This allows inferences about the reactivity of testing.

Instrumentation

In addition to participants changing from pretest to posttest, the data collection process may change. This change could be due to an alteration in the person or method for collecting data. This threat is especially salient when the data are collected through observational methods and the observer is replaced during the course of the study (Shaughnessy & Zechmeister, 1994). The new person assigned to observe may be less experienced than the previous observer or use slightly different criteria for judging and categorizing observations.

Changes in the method of collecting data may also alter the data obtained. For example, a researcher studying insurance use for psychiatric services relies on data collected by the insurance company. If the company changes to a new system with different coding structures, the data given to the researcher may be very different from the data obtained earlier. In this situation, findings of changes in service use may be due to actual changes or to alterations in the coding process. This is a common problem in using administrative data and requires a vigilant investigator. Instrumentation changes are also common in the study of young children. For tests to be age appropriate, the method for collecting data is often different. Measuring behaviors in an infant would be different from measuring behaviors in a 10-year-old child.

Statistical Regression to the Mean

If respondents are assigned to treatment or comparison groups on the basis of an extreme and unreliable pretest score, the study is at risk for the artifact of statistical regression to the mean. Statistical regression to the mean occurs when scores are inflated or deflated due to error in measurement. Because the pretest score may be abnormally high or low, subsequent testing will most likely yield scores closer to the population mean (Cook & Campbell, 1979). A pretest that measures depressive symptoms may question a woman who had very little sleep and was feeling a bit overwhelmed. Her depressive symptoms may appear quite high at that time. However, retesting after she has had a good night's sleep and a productive day at work may show very little depressive symptomology. Her pretest score was unexpectedly high, but later testing will be much closer to the average population score of symptoms. Regression to the mean is especially a problem when a group is selected or self-selects based on a high and unreliable score. Thus, selecting the sickest children for treatment will almost always result in improvement regardless of the treatment.

Statistical regression to the mean is associated with random measurement error, which is often due to unreliability of measures. Although the use of more reliable measures will not eliminate this problem, it can minimize the influence of statistical regression to the mean (Shadish et al., 2002).

Selection

How participants are selected can be a threat to validity when the average person in one of the groups differs from the average person in the other groups. This nonequivalence confounds treatment effects with participant characteristics, making causal inferences difficult. This threat is greatest in quasi-experimental research because random assignment is not employed. As described at the start of the chapter, the Fort Bragg evaluation has been the largest nonequivalent comparison group study in children's mental health services. Because random assignment was not used, the threat of selection bias is present. Several statistical techniques can be used to ameliorate selection factors (e.g., propensity analysis), but it is better, when feasible, to avoid selection factors through the use of a well-planned and implemented research design.

Attrition (Mortality)

Attrition, also known as mortality, has an effect on internal validity similar to that caused by selection. Attrition occurs when people drop out of a study and do not complete posttests. If the people who drop out of one group are different from those that remain in another group, individual participant characteristics can influence the outcomes. In such cases, causal inferences are difficult because one cannot determine if differences are due to the treatment or participant characteristics. Imagine a therapy targeted at reducing depression. What if some people in the treatment group begin to feel much better after therapy, decide they no longer need it, and stop attending sessions? This results in the treatment group containing only people who have not felt a large improvement. Comparisons of the therapy to no-treatment groups may show no or even negative effects, when the treatment may have been quite effective. Almost every mental health study has attrition. It is important that the investigator attempt to retain participants, record in detail the extent of attrition (Cordray & Pion, 1993), and conduct an attrition analysis to determine if there are biases (Foster & Bickman, 1996).

Demoralization and Compensatory Rivalry

In addition to participants being different from the way they were at the start of the study, they can also behave differently during the study. Participants who somehow learn that they are in the comparison group may feel disappointed about not being in the treatment group and simply give up. This demoralization will tend to inflate the differences between the groups because the comparison group will perform more poorly than a typical comparison group. Conversely, rather than give up, the comparison group may try harder to overcome differences from the group that receives treatment.

compensatory rivalry by the comparison group will reduce the difference between the groups, thus reducing the treatment effect. Whether participants know their group assignment is dependent on the research design and consent procedures of the study.

Contamination and Diffusion

Elements of the treatment may unintentionally be provided to the comparison group in two ways. First, participants in the comparison group may accidentally receive some of the treatment through communication with the treatment group. This is known as *diffusion of treatment.* Also, participants may accidentally get aspects of the treatment from program personnel. This is known as *contamination of the comparison group.* For example, in a program that provides hot meals to children in the treatment group, extra food may be given to the comparison group children as well. Both diffusion of treatment and contamination of the comparison group will minimize the difference between the two groups because to some degree they are both getting the treatment.

Interactions: Selection Maturation, Selection History, Selection Instrumentation

Although the threats to validity are described individually, they seldom work in isolation. Of the threats to internal validity, several are more likely to work in conjunction. These are (a) selection maturation, (b) selection history, and (c) selection instrumentation. Selection maturation is the additive effect of nonequivalence groups that are also maturing at differential rates. Selection history refers to nonequivalent groups that experience different influential historical events. Selection instrumentation occurs when nonequivalent groups are tested using different measurements. Measurement differences could include a change in measures, time intervals of measurement, or the presence of ceiling or floor effects in the data.

Ambiguity About the Direction of Causal Inference

As mentioned earlier, the cause must precede the effect in order for the researcher to determine causality. If measurement is cross-sectional, meaning it occurs only at one point in time, then it is difficult to claim that A causes B and B did not cause A, or perhaps that some other variable, such as C, caused A and B. When both A and B are measured at the same time, then the study is correlational because temporal order is unknown and no inferences about causality can be made.

Importance of Internal Validity

Traditionally, internal validity has been written about as the most important type of validity. It is the only type of validity that directly examines a causal

relationship. As Campbell and Stanley (1963) claimed, "Internal validity is the basic minimum without which any experiment is uninterpretable" (p. 5). Often, the more complicated the study question, the more frequently it will encounter potential threats to validity (Mark, 2000).

In addition to identifying threats to interpreting causal relationships, the researcher is interested in specifying the contingencies of such relationships. Examples of these contingencies include: "Under what conditions and to whom does the causal relationship apply?" Contingency questions are addressed through construct and external validity.

Construct Validity

Construct validity is focused on whether the study is measuring what it intends to measure. To make causal inferences, it is necessary to ensure that the study is assessing the intended cause and effect and has not mislabeled or confounded them with other variables during operationalization. The measurement of the cause and effect variables should have convergent and discriminant validity, meaning that it is similar to other measures of the same construct and unlike measures of different constructs. For example, a measure of major depression should be similar to some measures of bipolar disorder but different from measures of schizophrenia. In mental health services research, the construct validity of the cause (i.e., services or treatment) is usually very weak. There are no clear operational definitions of what is meant by such "treatments" as outpatient care or hospitalization. These can be seen as simply locations of treatment and not as describing treatment. Similar problems exist in the introduction of new treatments or services such as wrap-around. A clear and precise definition is required to enhance the construct validity of the service or treatment being delivered (i.e., cause). Additionally, a method for monitoring the fidelity of implementation of services or treatments is needed to ensure that the construct is actually delivered as planned and promised. Variability in implementation can also affect statistical conclusion validity. In children's mental health, we have less of a problem with the construct of effect because usually a great deal of effort is expended to determine the psychometric qualities of the outcome measures. However, even if these concerns are dealt with, there are still threats that the investigator must be vigilant about.

There are 10 potential threats to construct validity. Each of these threats can occur at the same time and greatly influence any interpretation that may be drawn from the data.

Inadequate Preoperational Explication of the Constructs

Prior to implementing a study, researchers must clearly articulate the constructs they intend to measure. They must clearly state what the cause and the effect are. Mislabeling them may lead to incorrect causal inferences.

Mono-Operational Bias

Most studies use only one measure of a construct. However, the use of only one measure will tend to inadequately represent the construct and may include irrelevancies and other constructs (Shadish et al., 2002). It is often not much of an added cost to include more measures of a construct.

Mono-Method Bias

In addition to using multiple measures of a construct, using different methods of data collection is preferable. This is due to the risk that the method of data collection will influence the data obtained. For example, if all the measures for self-efficacy are given as written surveys and some participants are uncomfortable reading, they may feel less efficacious while completing the form and report lower levels of self-efficacy. However, an interview with the same questions may yield different, more favorable, responses.

Hypothesis Guessing

Another threat to construct validity is that participants will be able to guess the hypothesis of the study and alter their behavior as a result. The best way to avoid this threat is to make the hypotheses difficult to identify, decrease the level of reactivity of the study, and, if possible, give different hypotheses to different participants (Cook & Campbell, 1979).

Evaluation Apprehension

Many people are apprehensive about being evaluated and may alter their behaviors to appear in a more positive light (Campbell & Russo, 1999). This evaluation apprehension typically results in people attempting to portray themselves as competent and psychologically healthy (Cook & Campbell, 1979). Such behaviors may magnify or obscure treatment effects.

Experimenter Expectancies

In addition to the participants having expectations about the purpose of the study, researchers might have expectations that could bias data. If the researcher believes in a treatment and is aware of who obtains the treatment there is an opportunity for conscious or unconscious bias. This problem can be avoided through the use of masking, in which the researcher is unaware of either the research hypothesis or the group membership of each participant. However, it is usually not possible to mask the person delivering the treatment. In psychotherapy research, this may lead to the allegiance effect. This effects occurs in the situation where positive results are only associated with the investigator who developed the therapy being tested.

Confounding of Constructs

Confounding of constructs occurs when the cause and effect vary at different levels such that A at one level causes B but at a different level has no effect on B. This is most common when continuous constructs are measured discretely or when only one level of a construct is measured. For example, a medication at 1 mg will reduce seizures, but at less than .8 mg or above 1.4 mg it will have no effect. A design that only measures 2 mg would be insensitive to the dose response and would erroneously indicate that the medication had no effect on seizures.

Interaction of Different Treatments

Studies that provide more than one treatment to the same participants may yield an effect from the interaction of the two treatments. Although this is more common in laboratory studies than field settings, there is a risk to testing more than one treatment at a time. To avoid this threat, researchers should test each treatment separately or provide several treatment groups, one for each treatment and one for each interaction. However, in services research it is often not feasible to isolate the critical elements of an intervention.

Interaction of Testing and Treatment

The treatment and testing process can potentially interact, reducing the generalizations that can be made about the effects of the treatment. For example, would a study that used three pretests have found the same effects if it had one pretest? The interaction of the treatment with the method and time interval of testing may have an effect on the conclusions that are drawn.

Restricted Generalization Across Related Constructs

If a causal relationship is established, it is necessary to determine the breadth of effects that may be influenced by the cause. For example, a program to reduce hyperactive behavior in children may show success in increasing overall attention but have little influence on other constructs, such as school performance and compliance with adult demands. How findings can generalize to other constructs is an important question for social interventions.

Overall, construct validity is important for applications of findings from studies. If a causal relationship is identified, it is necessary to determine that the constructs underlying the relationship are valid. For example, a researcher could conduct a study to improve mathematic ability through an intensive program that uses word problems. As a consequence of going over word problems, the participants improve their reading ability. At the conclusion of the study, the children perform better on mathematical word problems. Although a causal relationship was demonstrated, the cause was mislabeled. The improvement in the testing was not due to improved mathematical skills but increased reading ability. Inferences and applications from this study may be invalid while the causal relationship may have been internally valid.

ernal Validity

Social research is often focused on identifying causal relationships to determine how best to help people. The process of generalizing findings from a study to other people, settings, or times is referred to as *external validity* and is based on a correspondence between target populations, study populations, and the achieved sample. The more representative of the target population the people in the study are, the higher the external validity. Like the other forms of validity, external validity has threats as well. These are (a) an interaction of selection and treatment, (b) interaction of setting and treatment, and (c) interaction of history and treatment.

Interaction of Selection and Treatment

Selection bias affects not only causal inferences but also generalizations about findings. Perhaps the people who volunteer for a study are systematically different from those who do not. In such cases, selection is not a threat to internal validity because volunteers would be present in both the treatment and comparison groups; however, the interaction of selection and treatment could threaten external validity such that the findings of the study may not apply to other (nonvoluntary) people.

Research on the characteristics of people who volunteer for studies has shown them to be more motivated to comply (West & Sagarin, 2002), better educated, have higher incomes, and be of nonminority group status (Rosnow & Rosenthal, 1976) compared to the typical members of the population. If participants who volunteer for research are different from those who do not volunteer, the strength of generalizations of findings to other groups is weak.

Interaction of Setting and Treatment

The setting in which the treatment is implemented may also influence the generalizability of findings. For example, a program that is very effective in an in-service psychiatric hospital may be less effective when provided as an after-school program. Perhaps the intensive setting of a residential facility is more engaging for participants, whereas seeking services at school is perceived as stigmatizing. In this example, both the treatment and the setting interact with each other, influencing the application of causal relationships. The same program may be effective in one setting but not another. Of greater concern is the possibility that results obtained from a demonstration project cannot be replicated when applied in more typical settings.

Interaction of History and Treatment

The effects of treatment are influenced by the historical context in which they occur. A treatment may be very effective at one point in time but not very effective at another. In the area of mental health, attitudes and public awareness about mental illness have been changing. A program to identify children with

mental health problems may have experienced less success in the 1950s than it would today, when seeking services is more socially accepted. The program may be exactly the same in 1954 and 2004 but be successful only in the present day due to children and families' being more willing to seek services and comply with treatment.

RANDOMIZED EXPERIMENTS VERSUS QUASI-EXPERIMENTAL DESIGNS: TRADE-OFFS

Experimental and quasi-experimental designs are susceptible to threats to validity. Some designs are more vulnerable to certain threats than others; however, none are immune. Additionally, all four types of validity are interrelated. As noted by Cook and Campbell (1979), "increasing one kind of validity will probably decrease another kind" (p. 82).

Although randomized experiments greatly reduce the threat of selection bias and therefore increase internal validity, they may create a nontypical environment and therefore reduce external validity. For example, an experimental study may be able to randomly assign participants, but few programs have the resources or flexibility to implement such a procedure. Experimental studies may strengthen statistical conclusion validity through increasing participant numbers and number of measurement points, but this may be more burdensome to participants and result in greater attrition and more unreliability in measurement, thus reducing statistical and internal validity. Using multiple measures to improve construct validity may increase cost, resulting in lower numbers of participants, and may be more burdensome, leading to greater attrition. In this case, both internal validity and statistical conclusion validity will be lower.

All research designs involve trade-offs among the different types of validity. Therefore, researchers need to determine, a priori, which types of validity are most important. For instance, is the researcher predominately interested in determining the causal relationship between two variables or in the generalizability of the effects of a program to other groups? An explicit statement of the goals of the research will help determine which threats will be most severe and what design steps are needed to control for them.

Although random assignment helps protect against the threat of selection bias, randomized experiments are not free from its effects. If attrition is high or differential, the protective effects of randomization may be lost. Additionally, quasi-experimental designs that lack randomization can still control for its effects. Regression discontinuity designs enable researchers to determine the assignment criteria and control for it in subsequent analyses. Also, statistical procedures such as the use of propensity scores in nonequivalent comparison group designs allow researchers, through the use of logistic regression, to predict group membership based on scores on other relevant factors (Rosenbaum & Rubin, 1985; Rubin & Thomas, 2000).

Although randomized experiments are viewed as the gold standard for research, quasi-experimental designs, when designed well, can support the same causal inferences while increasing external validity. This chapter placed a great deal of emphasis on the types of and threats to the validity of a study. Armed with an understanding of these threats, researchers will be able to design better studies and consumers of research will be able to judge the quality and veracity of conclusions.

APPENDIX 4.A

Ethical Considerations for Research in Children's Mental Health Services

Awareness of the rights of people participating in research is relatively recent, with most legislative changes occurring since the 1960s. When considering the historical lack of protection of human participants, most people recall the atrocious medical experimentation conducted by the Nazis during World War II. However, even the United States has a very checkered past of infringing on the rights of its citizens when conducting medical research. The most common of our unprotected participants have been veterans, children, inmates, psychiatric patients, impoverished minorities, and persons with developmental disabilities.

As a result of public attention, especially regarding the Tuskegee Syphilis Study, the National Research Act of 1974 was passed. This act created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, which eventually published the Belmont report in 1979. The Belmont report identified three essential areas for ethical research: (a) respect for persons such that each person enters into research voluntarily and well informed; (b) beneficence as a rule so that research does not harm participants, maximizes their benefit, and minimizes harm to them; and (c) justice so that every participant is treated fairly and not exploited.

Additionally, the U.S. Department of Health and Human Services (DHHS) regulations for the protection of human participants (Title 45 *Code of Federal Regulations* Part 46) offer special protections for children and adolescents. In addition to consent from parents, researchers must explain the study to children and obtain their assent to participate. This requirement can be even more challenging for research on children's mental health services, in which the function and comprehension of some children, adolescents, and families may be lower than those of typically developing peers. Additionally, confidentiality is of utmost importance because of potential stigmatization and discrimination.

To ensure that the recommendations of the Belmont report and regulations of DHHS were used, many universities and research institutions created IRBs. These committees, typically composed of researchers, community members, and religious leaders, review proposed research projects and ensure that they meet federal and institutional criteria for ethical research. IRBs also conduct training for investigators and periodically audit research projects to ensure compliance with proposed protocol. Last, participants may contact the IRB if they suspect their rights as participants are not being protected. At research institutions, the IRB has the ability to halt research.

All research, especially research with children, should address participants' rights, protection, benefit, and potential risk. Ideally, an external body such as an IRB, will oversee empirical works to promote ethical behavior. Because children, especially children with special needs, are considered

vulnerable population, researchers must design consenting and assenting procedures that best inform them of their rights and ensure their confidentiality. Ethical research does not entail a static, rule-following process, but rather an iterative process of balancing research questions with the best interest of children, their families, and society. Although there may be an added challenge for research with children with mental disorders, ethics are a paramount concern in design and execution. For more information on ethical research in children's mental health, we highly recommend Hoagwood, Jensen, and Fisher (1996), *Ethical Issues in Mental Health Research with Children and Adolescents.* The protection of human participants applies equally to quasi-experimental and experimental designs. Both warrant the same level of ethical considerations and scrutiny in recruitment, implementation, and execution.

REFERENCES

Bickman, L. (1996a). A continuum of care: More is not always better. *American Psychologist, 51*(7), 689–701.

Bickman, L. (Ed.). (1996b). Special issue: The Fort Bragg experiment. *Journal of Mental Health Administration, 23*(1).

Bickman, L., Guthrie, P. R., Foster, E. M., Lambert, E. W., Summerfelt, W. T., Breda, C. S., et al. (1995). *Evaluating managed mental health services: The Fort Bragg experiment.* New York: Plenum Press.

Bickman, L., Smith, C. M., Lambert, E. W., & Andrade, A. R. (2003). Evaluation of a congressionally mandated wraparound demonstration. *Journal of Child and Family Studies, 12*(2), 135–156.

Bickman, L., Summerfelt, W. T., & Noser, K. (1997). Comparative outcomes of emotionally disturbed children and adolescents in a system of services and unusual care. *Psychiatric Services, 48*(12), 1543–1548.

Boruch, B. F. (1997). *Randomized experiments for planning and evaluation.* Thousand Oaks, CA: Sage.

Campbell, D. T., & Russo, M. J. (1999). *Social experimentation.* Thousand Oaks, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cohen, J., & Cohen, P. (1983). *Applied regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation.* Boston: Houghton Mifflin.

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi experimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, pp. 491–576). Palo Alto, CA: Consulting Psychologists Press.

Cordray, D. (2000). Enhancing the scope of experimental inquiry in intervention studies. *Crime and Delinquency, 46*(3), 401–424.

Cordray, D., & Pion, G. (1993). Psychosocial rehabilitation assessment: A broader perspective. In R. L. Glueckauf, L. B. Sechrest, G. R. Bond, & E. C. McDonel (Eds.), *Improving assessment in rehabilitation and health* (pp. 215–240). Newbury Park, CA: Sage.

Corrin, W., & Cook, T. (1998). Design elements of quasi-experiments. *Advances in Educational Productivity, 7,* 35–57.

Foster, M., & Bickman, L. (1996). An evaluator's guide to detecting attrition problems. *Evaluation Review, 20*(6), 695–723.

Hays, W. (1994). *Statistics.* Fort Worth, TX: Harcourt Brace College Publishers.

Hendrick, T. B., Bickman, L., & Rog, D. J. (1993). *Applied research design: A practical guide.* Thousand Oaks, CA: Sage.

Hill, J., Rubin, D., & Thomas, M. (2000). The design of the New York school choice scholarship program evaluation. In L. Bickman (Ed.), *Research design* (pp. 155–180). Thousand Oaks, CA: Sage.

Hoagwood, K., Jensen, P., & Fisher, C. (1996). *Ethical issues in mental health research with children and adolescents.* Mahwah, NJ: Erlbaum.

Holland, P. (1989). Comment: It's very clear. *Journal of the American Statistical Association, 84,* 875–877.

Keppel, G. (1991). *Design and analysis: A researcher's handbook.* Upper Saddle River, NJ: Prentice Hall.

Lambert, E. W., Wahler, R. G., Andrade, A. R., & Bickman, L. (2001). Looking for the disorder in conduct disorder. *Journal of Abnormal Psychology, 110*(1), 110–123.

Lipsey, M. (1990). *Design sensitivity: Statistical power for experimental research.* Thousand Oaks, CA: Sage.

Lipsey, M. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods.* Thousand Oaks, CA: Sage.

Lipsey, M. (2000). Statistical conclusion validity for intervention research: A significant ($p < .05$) problem. In L. Bickman (Ed.), *Validity and social experimentation* (pp. 101–120). Thousand Oaks, CA: Sage.

Lipsey, M., & Cordray, D. (2000). Evaluation methods for social intervention. *Annual Review of Psychology, 51,* 345–375.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mark, M. (2000). Realism, validity, and the experimenting society. In L. Bickman (Ed.), *Validity and social experimentation* (pp. 141–168). Thousand Oaks, CA: Sage.

McCleary, R. (2000). Evolution of the time series experiment. In L. Bickman (Ed.), *Research design* (pp. 215–234). Thousand Oaks, CA: Sage.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* Washington, DC: Belmont Conference Center at the Smithsonian Institution.

Reichardt, C. (1992). Estimating effects of community prevention trials: Alternative designs and methods. In H. D. Holder & J. M. Howard (Eds.), *Community prevention trials for alcohol problems: Methodological issues* (pp. 137–158). Westport, CT: Praeger.

Reichardt, C., & Mark, M. (1998). Quasi-experimentation. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods* (pp. 193–228). Thousand Oaks, CA: Sage.

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88,* 245–258.

Rosenbaum, P., & Rubin, D. (1985). Constructing a control group using multivariate matched sampling incorporating the propensity score. *American Statistician, 39,* 33–36.

Rosenthal, R., & Rosnow, R. (1991). *Essentials of behavioral research: Methods and data analysis.* Boston: McGraw-Hill.

Rosnow, R. L., & Rosenthal, R. (1976). The volunteer subject revisited. *Australian Journal of Psychology, 28*(2), 97–108.

Rubin, D., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton Mifflin.

Shaughnessy, J., & Zechmeister, E. (1994). *Research methods in psychology.* New York: McGraw-Hill.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Solomon, R. (1949). An extension of control group design. *Psychological Bulletin, 46,* 137–150.

West, S., & Sagarin, B. (2002). Participant selection and loss in randomized experiments. In L. Bickman (Ed.), *Contributions to research design: Donald Campbell's legacy* (pp. 117–154). Thousand Oaks, CA: Sage.

Wortman, P. (1994). Judging research quality. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 97–110). New York: Russell Sage Foundation.