"The world will not stop and think—it never does, it is not its way; its way is to generalize from a single sample."
—Mark Twain

5

Validity

What Makes a Study Strong?

CHAPTER OUTLINE

LEARNING OUTCOMES

1. Detect potential threats to statistical conclusion validity in published research.
2. In a given study, determine if the researcher adequately managed potential threats to statistical conclusion validity.
3. Detect potential threats to internal validity in published research.
4. In a given study, determine if the researcher adequately managed potential threats to internal validity.
5. Detect potential threats to external validity in published research.
6. In a given study, determine if the researcher adequately managed potential threats to external validity.

KEY TERMS

alternative treatment threat
assignment threat
attrition
Bonferroni correction
compensatory demoralization
compensatory equalization of treatments
convenience sampling
covary
ecological validity
effectiveness study
efficacy study
experimenter bias
external validity
fishing
Hawthorne effect
history threat
instrumentation threat
internal validity
matching
maturation threat
mortality
order effect
participant bias
power
practice effect
Pygmalion effect
random assignment
random sampling
regression to the mean
replication
response rate
Rosenthal effect
sampling error
selection threat
statistical conclusion validity
testing effect
validity

Introduction

The purpose of this chapter is to make sure that you *don't* do what Mark Twain suggests in the opening quote and generalize from single research samples. You will learn how to stop and think about data both from a single sample and multiple samples, with the goal of using the information in evidence-based practice.

When evaluating the strength of evidence, there are certain axioms that practitioners tend to rely on, such as "Randomized controlled trials are the strongest design" and "Large sample sizes provide more reliable results." Although true in many cases, there can be exceptions. To be a critical consumer of research, it is essential to understand the *why*s behind these assertions. Why is a large sample size desirable? Why are protections inherent in randomized controlled trials? Even with a large-sample, randomized, controlled trial, other factors may compromise the validity of the study.

This chapter explains the concept of validity, describes different threats to validity, and identifies possible solutions to these threats. This information will increase your ability to critically appraise research. If you have a good grasp of the possible threats to validity and the ways in which these threats can be managed, you will be able to evaluate the strength of evidence and become an evidence-based professional.

Validity

When thinking about the validity of a study, consider the terms *truthfulness*, *soundness*, and *accuracy*. **Validity** is an ideal in research that guides the design, implementation, and interpretation of a study. The validity of a study is enhanced when sound methods allow the consumer to feel confident in the findings. The validity of a study is supported when the conclusions drawn are based on accurate interpretations of the statistics and not confounded with alternative explanations. The inferences that are drawn from a study will have greater validity if they are believable and reflect the truth. This chapter describes three types of research validity: (1) statistical conclusion validity, (2) internal validity, and (3) external validity. Chapter 6 addresses a different type of validity concerned with assessments used by researchers.

Statistical Conclusion Validity

Statistical conclusion validity refers to the accuracy of the conclusions drawn from the statistical analysis of a study. Recall that with most inferential statistics, a p value is calculated; conventionally, if the p value is < 0.05, the conclusion is one of statistical significance (i.e., there is a statistically significant difference or there is a statistically significant relationship). As an evidence-based practitioner, there may be reasons why you should question the researchers' conclusions that are presented in a research article.

Threats to Statistical Conclusion Validity

In Chapter 3, mistaken statistical conclusions were described in terms of Type I and Type II errors. As an evidence-based practitioner, you can identify potential errors by increasing your awareness of research practices that lead to error. Specific threats to statistical conclusion validity, their relationship to error type, and methods researchers use to protect research from those threats are described in this chapter. Table 5-1 outlines the threats to statistical conclusion validity, confounding factors that interfere with statistical conclusion, and methods for protecting against these threats.

TABLE 5-1 Threats to Statistical Conclusion Validity and Their Protections

Threat	Type of Error	Confounding Factor That Interferes With Statistical Conclusion	Protection
Fishing	Type I	• Researcher searches data for interesting findings that go beyond the initial hypotheses. • Conclusions may be due to chance.	• Use statistical methods that adjust for multiple analyses. • Conduct a second study to test the new hypothesis with different participants.
Low power	Type II	• A difference or relationship exists, but there is not enough statistical power to detect it.	• Increase alpha level. • Ensure that intervention is adequately administered to obtain optimal effect size. • Increase sample size.

Fishing

Fishing is a euphemism that refers to looking for findings that the researcher did not originally plan to explore. Ideally, when a researcher conducts a study, a hypothesis is developed before collecting data. Once the data are collected, a statistical analysis is applied to test the hypothesis. However, not infrequently, researchers will explore existing data in what is sometimes called a "fishing expedition" or "mining for data." In other words, the researcher is letting the data lead the way toward interesting findings. Although there are legitimate reasons for delving into the data, the risk in fishing is that the researcher will see interesting differences or relationships that may not be true and instead are due only to chance. In other words, the researcher has committed a Type I error by finding a difference that does not exist.

Typically many analyses are conducted in a researcher's search for findings. Previously, you learned that when alpha is set at 0.05, the researcher is willing to take a 5% risk that the difference or relationship is not true but is due to chance. However, this applies only to a single analysis. Each time another analysis is performed, there is a greater risk that the finding is due to chance. Researchers often explore their data for unexpected findings, which can lead to important discoveries. However, protections should be in place so that chance findings are not misleading.

Protection Against Fishing Threats

As an evidence-based practitioner, you may suspect that a fishing expedition has occurred when the results of the study are not presented in terms of answers to a research hypothesis. A straightforward researcher may acknowledge the exploration and, if it is a robust study, will describe how threats to Type I error were addressed.

One way that researchers can protect against fishing threats is to use statistical procedures that take into account multiple analyses. There are many such procedures, but the simplest one conceptually is the **Bonferroni correction**. With the Bonferroni correction, the alpha level of 0.05 is adjusted by dividing it by the number of comparisons. For example, if six comparisons were made, 0.05/6 = 0.0083, meaning that the acceptable alpha rate is much lower and much more conservative than the initial 0.05.

Another method that protects against fishing threats involves conducting another study to test the new hypothesis discovered when the data were explored. For example, consider a researcher who tested a new intervention and found that the initial analyses did not show the intervention to be more effective than the control. However, upon deeper analysis, the researcher discovered that men experienced a significant benefit, whereas women stayed the same. A new study could be conducted to test this hypothesis. If the second study resulted in the same findings, there would be stronger evidence to conclude that only men benefit from the intervention.

Low Power

Power is the ability of a study to detect a difference or relationship. Power is based on three things: sample size, effect size, and alpha level. The larger the sample is, the more powerful the study is. It is easier to detect a difference when you have many participants. Likewise, if you have a large effect, you will have greater power. If an intervention makes a major difference in the outcome,

it will be easier to detect that difference than if an intervention makes only a minor difference.

Recall from Chapter 3 that a Type II error occurs when no difference is found, but in actuality a difference is present. This occurs because of low power and is most often the result of small sample size. When you review a study with a small sample size that does not find a difference or a relationship, low power is a potential threat to statistical conclusion validity. However, it is also possible that, even with a large sample, the researcher does not find a difference or relationship.

Protection Against Low Power Threats

Power can be increased by changes in the alpha level, effect size, or sample size. In exploratory analyses, the researcher may utilize a higher alpha level, such as 0.10 instead of 0.05; however, in doing so, the researcher takes a greater chance of making a Type I error. It is more difficult to change the effect size, but the researcher needs to ensure that everything is in place to test whether the intervention is effective (e.g., trained individuals administer the intervention, strategies that foster adherence are used, etc.).

The simplest way to increase the power of a test is to increase sample size. However, it can be costly in terms of both time and resources to conduct a study with a large sample. Researchers often conduct a power analysis to determine the smallest sample possible to detect an effect given a set alpha level and estimated effect size.

The potential for Type II errors provides a strong rationale for using large samples in studies. With a large sample, a researcher is unlikely to make a Type II error. However, there are additional benefits to having a large sample size. With a large sample, outliers are less likely to skew the results of a study. For example, consider the average of the following six scores on an assessment:

$$5 + 4 + 5 + 3 + 26 + 5 = 48/6 = 8$$

The score of 26 is an outlier, when considering the other scores. The mean score for this sample of 6 is 8. When you look at each individual participant, 8 is a much higher score than the majority of the participants received. A single outlier misrepresents the group as a whole.

Now consider a sample of 40 participants:

$$5 + 4 + 5 + 3 + 26 + 4 + 3 + 4 + 4 + 5 + 4 + 5 + 5 + 4 + 5 + 3 + 3 + 4 + 3 + 4 + 4 + 5 + 4 + 5 + 5 + 4 + 5 + 3 + 4 + 3 + 3 + 4 + 5 + 4 + 3 + 5 + 4 + 3 + 3 + 4 = 183/40 = 4.58$$

The outlier has a weaker effect on the group as a whole, and the mean for this sample is more in line with the typical scores.

Another benefit of a large sample is that, the larger the sample, the more likely it is that the sample will represent the population. This fact is particularly relevant for survey research. Not only will a large sample represent the population, but it shows that more individuals are likely to respond when invited to complete the survey. The **response rate** is the number of individuals who respond to a request to participate in a survey or other research endeavor. The larger the response, the more accurate the results. In the case of survey research, the response rate is determined by dividing the number of surveys that were completed by the number of surveys that were administered. For example, if 200 surveys were sent out, and 150 people completed and returned them, the response rate would be 150/200 = 75%.

Individuals who choose not to participate in a study may opt out of the study for a particular reason and, in doing so, bias the results. For example, if you are conducting a satisfaction survey for your therapy program and only 25% of your clients respond, it is possible that the individuals who responded are either highly dissatisfied or highly satisfied and therefore more motivated to voice their opinions.

Exercise 5-1

Identifying Threats to Statistical Conclusion Validity (LO1 and LO2)

Read the following scenario and identify which practices present potential threats to statistical conclusion validity. Suggest methods for controlling these threats.

A new researcher who is a therapist wants to collect data to examine the efficacy of three different orthoses for a specific hand condition. The researcher plans to recruit clients from her clinic and expects that approximately 10 individuals will have the hand condition of interest. The following outcomes will be measured: pain, range of motion, fine motor control, and function. The researcher has no expectation as to which orthosis will provide the better outcome.

QUESTIONS

1. Why are there fishing threats?

2. How could the researcher address the fishing threats?

3. Why are there threats due to low power?

4. How could the researcher address the low power threats?

Internal Validity

When evaluating a study for evidence, it is necessary to consider internal validity and how it may affect the study outcomes. A study has **internal validity** when the conclusions drawn from the results are accurate and true. Validity is not an either/or situation, but rather a matter of degree. For example, a study that examines the effectiveness of social stories in children with autism concludes that children in the intervention group had greater improvement in their social skills than children in the control group. If the study were internally valid, this would mean that it was truly the social stories intervention that improved the social skills. However, there is always the possibility that there is an alternative explanation for the study results. Perhaps the difference was due to chance. Or it could be that the attention the children received is what made the difference, and not the intervention itself. Perhaps the individuals who administered the outcome assessments were biased and tended to give higher scores to the individuals in the intervention group. Although you can never be certain that the results of a study are entirely accurate, certain features of the study can greatly increase your confidence in its accuracy and validity.

Threats to Internal Validity

When examining internal validity, ask yourself, "Is there an alternative explanation for these study results?" Alternative explanations are often referred to as "threats" to internal validity. This section of the chapter characterizes common threats to internal validity, describes protections or solutions to avoid or minimize those threats, and identifies types of research situations in which these threats are most likely to occur. Table 5-2 summarizes the threats to internal validity and their protections.

Assignment and Selection Threats

Threats to internal validity can occur when a bias is present during the process of assigning or selecting

TABLE 5-2 Threats to Internal Validity and Their Protections

Threat	Confounding Factor Affecting Outcome/Alternative Explanation	Protection
Maturation	• Changes occur over time in participants as a result of development or healing.	• Use control groups. • Ensure baseline equivalence through random assignment or participant matching.
Assignment/ Selection	• Groups are not equal on some important characteristics.	• Random assignment • Participant matching • Statistical procedures such as covariance
History	• Events occur between the pretest and posttest.	• Use control groups. • Ensure short time between pretest and posttest. • Ensure protection against exposure to alternative therapies.

Continued

TABLE 5-2 Threats to Internal Validity and Their Protections *(continued)*

Threat	Confounding Factor Affecting Outcome/Alternative Explanation	Protection
Regression to the mean	• Extreme scores change and move toward the mean with repeated testing.	• Use control groups. • Exclude outliers. • Take the average of multiple measurements.
Testing/practice/ order effects	• Performance on measures changes due to exposure or some other feature of the testing experience.	• Use control groups. • Use measures with good test/retest reliability • Use alternate forms of measures. • Counterbalance the order of measures. • Take breaks if fatigue is anticipated.
Instrumentation	• Invalid or unreliable measures, tester error, or poor condition of the instrument result in inaccurate outcomes.	• Use measures with good reliability and validity. • Use measures that are sensitive to change. • Train the testers. • Maintain the instruments. • Blind the tester.
Participant and Experimenter Bias Threats		
Rosenthal/ Pygmalion effect	• Intervention leaders expect participants to improve, or participants expect to improve. • When a control group is involved, the expectation is that the experimental group will perform better than the control group.	• Blind the intervention leaders and/or participants. • Limit contact between intervention and control leaders and participants. • Establish and follow protocols for interventions.
Compensatory equalization of treatment	• Intervention leader's behavior encourages participants in the control group to improve to equal the intervention group, or control group participants are motivated to compete with the intervention group.	• Blind intervention leaders and/or participants. • Limit contact between intervention and control leaders and participants.
Compensatory demoralization	• Intervention leader's behavior discourages the control group, or participants are discouraged because of being in the control group.	• Blind intervention leaders and/or participants. • Limit contact between intervention and control leaders and participants.

Continued

TABLE 5-2 Threats to Internal Validity and Their Protections *(continued)*

Threat	Confounding Factor Affecting Outcome/Alternative Explanation	Protection
Hawthorne effect	• Participants improve because of attention received from being in a study.	• Blind intervention leaders and/or participants. • Ensure equal attention for intervention and control groups.
Attrition/Mortality	• Participants who drop out affect the equalization of the group or other characteristics of participants.	• Employ strategies to encourage attendance or participation. • Use statistical estimation for missing data. • Perform intent to treat analysis.

participants for groups. If there are differences between the groups on a baseline characteristic, this difference might affect the outcomes of the study.

Demographic characteristics, illness/condition issues, medication, and baseline scores on the outcome measures are examples of variables that should be considered when examining assignment and selection threats, because these characteristics can account for differences in outcome. An **assignment threat** indicates that the bias occurred when groups were assigned, whereas a **selection threat** indicates the bias occurred during selection—either the selection of participants or the selection of sites. For example, Killen, Fortmann, Newman, and Varady (1990) found that men responded better than women to a particular smoking cessation program. If at baseline there were more men in the intervention group and more women in the control group, the study would be biased toward finding more positive results than would exist with equal distributions of gender across the two groups. In another example, a splinting study for cerebral palsy may find that one group is receiving more antispasticity medication than another. The medication could then account for some or all of the differences in the outcomes.

Comparisons of the intervention and control groups at baseline should be provided in the initial section of the results section of a study so the reader can determine if there are differences between the groups. A table is typically provided that outlines the comparison of the groups on important demographic characteristics as well as the baseline scores of the outcome variables. An example of this type of comparison is shown in **From the Evidence 5-1**. The table comes from an intervention study examining the efficacy of an education and exercise program to reduce the chronicity of low back pain. The table compares the intervention and control groups at baseline (del Pozo-Cruz et al, 2012).

Protection Against Assignment Threats

Random assignment is the primary protection method used by researchers against assignment threats. In random assignment to groups, each research participant has an equal chance of being assigned to the available groups for an intervention or control. There are times when researchers are reluctant to use random assignment to a no-intervention control group, because it may be considered unethical to withhold treatment from a group. Sometimes this concern is managed by using a wait-list control group. The control group eventually receives the intervention, but not until the intervention group has completed the treatment.

Random assignment does not ensure equal distribution, which is particularly true with small samples in which extremes in a few individuals can greatly influence the group results. However, it works particularly well with larger samples because you are more likely to form equivalent groups. Therefore, when evaluating evidence, examine the group comparisons presented in the results section of the study.

Sometimes researchers use strategies such as **matching** study participants to ensure equal distribution on a particularly important characteristic. For example, if the researcher knows that the outcomes are likely to be influenced by a characteristic such as level of education, symptom severity, or medication, potential participants are identified and matched on the variable of interest, with one randomly assigned to the intervention group and one randomly assigned to the control group.

Another procedure that can be used to minimize assignment threats is statistical equalization of groups. If

FROM THE EVIDENCE 5-1

Table Comparing Study Groups

del Pozo-Cruz, B., Parraca, J. A., del Pozo-Cruz, J., Adsuar, J. C., Hill, J., & Gusi, N. (2012). An occupational, Internet-based intervention to prevent chronicity in subacute lower back pain: A randomized controlled trial. Journal of Rehabilitation Medicine, *44(7), 581–587. doi:10.2340/16501977-0988.*

Table I. Baseline Characteristics of Participants in the Study (n = 90)

Group	Control group (n = 44)	Intervention group (n = 46)	p
	Mean (SD)	Mean (SD)	
Age (years)	45.50 (7.02)	46.83 (9.13)	0.44
Sex (%)			
Male	11.4	15.2	
Female	88.6	84.8	0.59
Smokers, yes/no, %	50/50	56.5/43.5	0.53
Roland Morris Questionnaire score, points	11.65 (2.14)	12.28 (2.63)	0.22
TTO, points	0.78 (0.08)	0.75 (0.11)	0.23
SBST total score, points	4.38 (1.67)	4.36 (1.28)	0.95
SBST psychological score, points	2.36 (1.03)	2.28 (0.98)	0.70

p-values from t-test for independent measures or 2 test.
TTO: Time Trade Off; SBT: STarT Back Tool; SD: standard deviation.

Note A: The SBST is an outcome measure for the study.

Note B: The *p* value is above 0.05 for all comparisons, indicating that the two groups are comparable (i.e., there are no statistically significant differences) at baseline. This is particularly true of the SBST total score.

FTE 5-1 Question 1 *Are the two groups equivalent on all key characteristics—both demographic variables and outcome variables? How do you know?*

one group is older than another group, age can be **covaried** in the statistical analysis so that it does not influence the outcomes. If, when reading the initial section of the results section, you find that the groups are not equal on one or more important characteristics (which sometimes occurs, even with random assignment), check to see if the researcher handled this by covarying that variable, or at least acknowledging the difference in the limitations section of the discussion.

Maturation Threats

Maturation is a potential threat in intervention research involving health-care practitioners. *Maturation* refers to changes that occur over time in research participants. Two major types of **maturation threats** are particularly common in health-care research: (1) changes that occur as part of the natural growth process, which is particularly relevant for research with children; and (2) changes that occur as a result of the natural healing process, which is particularly relevant for research related to diseases and conditions in which recovery is expected. In other words, is it possible that if left alone the research participants would have changed on their own? Maturation is of greatest concern when the time period between the pretest and posttest is prolonged, such as during longitudinal studies or studies with long-term follow-up.

To illustrate the maturation threat, consider a study that examines an intervention for children with language delays. The study finds an improvement in language from

the pretest to the posttest; however, without adequate protection from other influences, it is difficult to determine whether the intervention caused the improvement or the change occurred as a result of developmental changes in language. Maturation would be an even greater concern if the study extended over a significant period of time, such as throughout a school year. Similarly, an intervention study examining changes in mobility for individuals after hip replacement would need to take into account the maturation threat, because individuals can experience improved mobility without therapy.

Maturation is in play whether the natural changes are positive or negative. When conditions result in a natural decline, the goal of therapy is often to reduce the speed with which that decline occurs. For example, if a therapist is using a cognitive intervention for individuals with Alzheimer's disease, it would be challenging to determine if a decline were less severe than would have occurred naturally over the course of the illness. However, studies can be designed with the proper protections to determine whether a particular intervention reduces the natural course of a decline in functioning.

Protections Against Maturation Threats

The primary protection against maturation threats is use of a control group. If the intervention group improves more than the control group, the difference between the two groups is more likely to be due to the intervention, even if both groups improve over time. The degree of improvement that the intervention group makes above and beyond the control group is likely due to the intervention.

Another protection against maturation threats is outcome scores that are similar at baseline for the control and intervention groups. This allows you to be more confident that the groups start out at a similar place, and makes interpretations of changes from pretest to posttest more straightforward.

Random assignment and matching of participants are additional strategies that increase the likelihood that the groups will be equal at baseline. (Random assignment and matching are described in detail in the section on assignment and selection threats.) In the results section of a research study, typically the first report of results is the comparison of the intervention and control groups; this includes pretest scores on the outcomes of interest and demographic variables that could affect the findings. Finally, you can be more certain that maturation is not a factor when the time between the pretest and posttest is short and when it is unlikely that changes would occur without an intervention.

History Threats

A **history threat** involves changes in the outcome or dependent variable due to events that occur between the pretest and posttest, such as a participant receiving an unanticipated treatment or exposure to an activity that affects the study outcome. In this case, the threat may also be referred to as an **alternative treatment threat**. For example, participants in a fall prevention program may start attending a new senior center that provides exercise classes with an emphasis on strength and balance.

In fact, any external event that can affect the dependent variable is a potential threat. A new teacher in a classroom who uses an innovative approach, participation in a health survey that draws attention to particular health practices, or a new fitness center opening in the participants' neighborhood could pose a threat to internal validity.

History can also have a negative effect on outcomes. A snowstorm might affect attendance, or scheduling a weight-loss program around the Thanksgiving and Christmas holidays could interfere with desired outcomes and act as a threat to internal validity.

Protections Against History Threats

History threats are avoided by many of the same strategies that are used to protect against maturation effects. The use of a control group provides protection, as long as both groups have the same potential exposure to the historical

FROM THE EVIDENCE 5-1 *(CONT.)*

FTE 5-1 Question 2 *Using this example, why is it important that participants in the intervention and control groups have similar scores at baseline on the STarT Back Tool? How does equivalence at baseline protect against maturation threats?*

event. Likewise, history is reduced as a threat when there is a shorter time between pretest and posttest. Researchers can also put protections in place to reduce exposure to alternative treatments, such as requiring participants to avoid alternative exercise programs or drug therapies while involved in the study. The researcher can include questionnaires or observations to help determine if events occurred that might affect the outcome.

Regression to the Mean Threats

Regression to the mean refers to a phenomenon in which extreme scores are likely to move toward the average when a second measurement is taken; extremely high scores will become lower, and extremely low scores will become higher. When taking a test for a second time, it is always possible—even likely—that you will not receive the exact same score. This phenomenon is especially predictable in individuals who initially score at the extremes of the distribution. At the ends of the distribution, it is less likely that a second test score will become even more extreme; instead, extreme scores tend to regress toward the mean. The "*Sports Illustrated* curse" serves as a case in point. It is often observed that after someone is featured in *Sports Illustrated*, that individual has a decline in performance. Regression to the mean would explain this observation, because the individual athlete is likely featured when he or she is at a peak of performance and superior to most if not all other athletes in that sport. Consequently, subsequent performance is likely to move toward the average, rather than improve.

In health-care research, study participants often start with extreme scores because of their condition. Therefore, when extreme scores are involved, regression to the mean should be considered a potential threat. **Figure 5-1** depicts the normal curve and illustrates the propensity for extreme scores to regress toward the mean; the extreme scores toward both ends of the continuum move toward the middle.

Protection Against Regression to the Mean Threats

Similar to history and maturation threats, regression to the mean is protected against by use of a control group. Once again, if the treatment group outperforms the control group, the difference between the groups is most likely due to the intervention. The importance of a control group should be more and more apparent; control groups are valuable because they address multiple threats to validity.

One other option is for researchers to exclude outliers from a study, although this tactic is not feasible when large numbers of participants could be classified as outliers. When small samples are necessary, the threat posed by an extreme score at baseline may be reduced by taking multiple pretest measures and using the average. For example, waist circumference can be challenging to measure accurately, so during testing, three measures may be taken and then averaged.

Testing Threats

Testing as an internal validity threat occurs when changes in test performance are a result of the testing experience itself. A **testing effect** is present when an earlier experience somehow affects a later testing experience. There are many different ways in which this can occur.

The testing experience often sensitizes participants to a desirable outcome. For example, the pretest may ask questions about following a home program, so the individual becomes sensitized to this behavior and begins following the program (as a result of the test, not the intervention). In another example, pedometers and other devices are often used as a measure of physical activity. The simple act of wearing the pedometer can influence how far an individual walks because the presence of the pedometer motivates the person to walk more, especially when the participant can see the readings. In this case, it is the wearing of the pedometer and not the intervention that causes the change. The tester can also influence the outcomes of the testing with behaviors such as providing cues to enhance performance, such as, "Try harder, you can do a few more."

Practice effects are a type of testing threat that occurs when exposure to the pretest allows the individual to perform better on the posttest. Prior exposure can mean the test is more familiar, the participant is

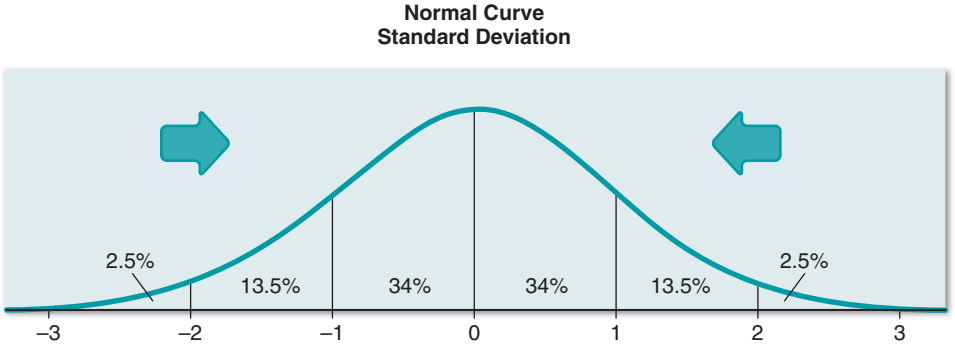

FIGURE 5-1 Normal curve, showing standard deviations and propensity for extreme scores to regress toward the mean. The shape of the curve suggests that individuals who score toward the ends are more likely to move toward the middle on a second testing.

less anxious, and the participant can adopt a strategy for improved performance at posttest. For example, students who receive a handwriting test before and after a handwriting intervention may do better on the second test, simply because of exposure and practice from the first test.

In the case of **order effects**, there is a change in performance based on the order in which the tests are presented. For example, in a long testing session there may be a decline in performance due to fatigue.

Protection Against Testing Threats

A measure with strong test-retest reliability is a good start for protecting against testing threats. Standardization, scripts, and training for the tester also can reduce biases that the tester may introduce. Instead of logs or diaries that can act as prompts to engage in a behavior, time sampling can be used to protect against testing threats. With time sampling, individuals receive a prompt, but the prompt is unexpected and random, and the individual or an instrument records what the person is doing at that point in time. Control groups are also beneficial with testing threats. If both groups receive some benefit from exposure to the testing situation, the difference between the control and intervention groups still represents the intervention effect.

Some measures are more vulnerable to practice effects than others. For example, a list learning test that assesses memory is less reliable the second time it is administered because the participant may remember words from the first time. In contrast, range-of-motion testing is less amenable to practice effects. Alternate forms are often used when a measure can be learned, such as a list of words. In this case alternate lists are made available, so that the individual is tested with a different set of words.

Instrumentation Threats

Instrumentation threats occur when a measure itself or the individual administering the measure is unreliable or invalid. Instrumentation threats are a common problem in research. A well-designed study is rendered useless if instrumentation poses a threat to its validity. When using mechanical or electronic measures, the quality, condition, and calibration of the instruments can affect outcomes. When an instrument is in poor condition, the measurements may be inaccurate. For example, it is recommended that the Jamar dynamometer be professionally calibrated on a yearly basis (Sammons Preston, n.d.).

Human error can also play a role in instrumentation threats. For example, a tester may provide incorrect instructions to a participant or make poor judgments when scoring an observational measure. The test itself may be a poor choice for the study, such as when it does not accurately measure the intended outcome; this represents an issue with validity of the test, which is a very important feature for an intervention study. For example, many therapy studies use self-report measures to assess an individual's functional performance. Although these measures are easy to administer, they may not provide accurate results. Sabbag et al (2012) found that performance measures were more accurate in assessing daily living skills in individuals with schizophrenia than was self-report.

Another aspect of instrumentation threat is the "ceiling effects" or "floor effects" of a measure. If ceiling effects are in play, the participants may have such a high score at the beginning that there is no room for improvement. Alternatively, the test may be so difficult (floor effects) that it is unlikely the researcher will be able to detect a significant change.

Protection Against Instrumentation Threats

Proper selection of measures is an essential protection against instrumentation threats prior to subject selection. It is essential to know the reliability and validity of measures used in a study and their sensitivity to change. Godi et al (2013) compared two balance measures—the Mini-BESTest and the Berg Balance Scale—in terms of their sensitivity in detecting change. They found that the Berg Balance Scale had greater ceiling effects, suggesting that the Mini-BESTest may be the better instrument to use in a study examining the efficacy of an intervention to improve balance.

Training of testers also provides protection against instrumentation threats. If multiple testers are used, inter-rater reliability among the testers should be established. Electronic and mechanical measures should receive the necessary maintenance and calibration. For example, audiologists are particularly cognizant of the importance of calibration, as testing of hearing impairment would be significantly compromised with a poorly calibrated instrument.

Experimenter and Participant Bias Threats

Experimenter bias is introduced when the research process itself affects the outcomes of the study, whether intentionally or unintentionally. Experimenter bias can be introduced by the person(s) providing the intervention.

A classic experimenter bias, known as the **Rosenthal effect** or the **Pygmalion effect**, occurs when the researcher sets up different expectations for the intervention and control groups. The term *Rosenthal effect* comes from an experiment by Rosenthal and Jacobson (1968) that involved teachers. Rosenthal communicated to some teachers that they should expect a strong growth spurt in intellectual ability from the students, whereas other teachers were not given this information. Students performed better when the teachers expected them to perform better. This study was set up to study this effect, but the same phenomenon can occur unintentionally when an intervention leader communicates an expectation of better outcomes from the intervention group. The higher expectations become a

self-fulfilling prophecy. Perhaps the leader provides more attention or enthusiasm, or works harder at providing the intervention, or the participants pick up on the leader's expectations and respond in kind.

Just being assigned to a particular group can introduce an experimenter bias. For example, without the leader's prompting, the control participants may want to compensate for not being picked for the intervention. In many rehabilitation studies, the control group receives standard treatment or "treatment as usual." If the control group is aware that the intervention group is receiving something new, they may try to compensate for this difference.

Another bias that can be introduced by the experimenter is **compensatory equalization of treatments**. In this case, the intervention leaders for the control group may feel compelled to work harder to compensate for the fact that the control group is not receiving the intervention. This type of bias is similar to the Rosenthal effect, but directed toward the control group. The control group may also respond in the other direction and feel discouraged because they are not receiving the intervention. In response, they may not try as hard or give up. This threat to validity is called **compensatory demoralization**. The threats of compensatory equalization and demoralization are more likely to occur when the leaders and/or participants of the control and treatment groups interact with one another.

Participant bias threats come into play when the participant's involvement in the study affects the outcomes. The **Hawthorne effect** occurs when participants respond to the fact that they are participating in a study and not the actual intervention (Mayo, 1949). The term comes from research conducted at the Hawthorne electric plant. Many variables were studied to determine what factors might affect productivity, such as lighting or changes in workstations. No matter what was studied and how insignificant the change, there was a change in productivity. It was concluded that the change and not the actual condition was resulting in greater productivity. The Hawthorne effect may occur because participants behave as expected or want to please the researcher.

In an interesting study examining the efficacy of ginkgo biloba in Alzheimer's disease, McCarney and colleagues (2007) examined follow-up as a confounding variable influenced by the Hawthorne effect. Some participants had minimal follow-up, whereas others had intensive follow-up. In other words, the intensive follow-up group received more attention from the researchers. The results indicated that participants receiving intensive follow-up had greater cognitive improvement than participants receiving minimal follow-up. This result is a particularly remarkable example of the Hawthorne effect, given that cognition was measured using an objective, standardized assessment (ADAS-cog) that included 11 cognitive tasks, such as word recall, orientation, and the ability to follow commands.

Protection Against Experimenter and Participant Bias Threats

Unlike many of the previous threats to validity, random assignment to an intervention and control group does not protect against experiment and participant bias. However, blinding of the intervention leaders and participants provides a strong protection against these threats. If the leaders and participants do not know whether they are providing or receiving the intervention, it is more difficult to introduce a bias. This is a common approach in drug trials in which a placebo is used in place of the actual medication.

In rehabilitation research, it is often difficult to blind intervention leaders and participants; therapists will know

Evidence in the Real World

How Lack of Blinding Participants Can Lead to Compensatory Equalization

In rehabilitation and therapy practices, it is difficult and in many cases impossible to blind the intervention leaders and participants to which group is receiving the experimental intervention. If you are the intervention leader, you have to know what you are leading (as opposed to offering a placebo pill); if you are a participant, you will likely know what intervention you are participating in.

In a real-life example, I was administering a weight-loss program for individuals with psychiatric disabilities. Participants were randomly assigned to either an intervention or a no-treatment control group. After the informed consent process, participants were told which groups they were assigned to. Some control participants voiced a desire to show the researchers that they could lose weight on their own, apparently compensating for not being assigned to the intervention. In some cases it worked, and indeed several control participants were successful in losing weight during the time they participated in the study.

To control for this confound, it would have been helpful for both groups to receive some form of intervention, so that neither group felt compelled to prove something based on their group assignment. As a reader of evidence, it is often hard to discern when individuals are responding to experimenter or participant bias; however, it is useful to know that the protections discussed in this section are in place to protect against potential problems.

they are providing an intervention and will typically know when they are providing the experimental intervention. A form of a placebo is provided in some rehabilitation studies when the control group receives an intervention that equalizes attention. When a new intervention is compared with standard therapy and when both groups receive the same amount of intervention time, experimenter and participant bias is less of a concern. Therefore, equal attention to groups is generally preferable to a no-treatment control. However, participants may know through the informed consent process that they are receiving the experimental intervention. Typically when individuals volunteer to be in a study, they want to receive the new intervention, so this can lead to disappointment if they are not assigned to the experimental condition.

Other methods can be used to minimize experimenter and participant bias. Clear protocols for the administration of the interventions can reduce bias. In some cases the therapists may be expected to follow the protocol and ideally do not know which intervention is expected to yield superior results. It is also helpful to limit interactions between intervention leaders to further prevent the development of bias. Similarly, keeping the participants in the intervention and control groups separate diminishes bias on the part of those receiving the treatment. In addition, it is often possible to blind the individuals who administer the outcome assessments in order to reduce or eliminate bias in scoring the assessments.

Attrition/Mortality Threats

Threats due to **attrition**, also called **mortality**, involve the withdrawal or loss of participants during the course of the study. The process of informed consent acknowledges that participants can withdraw from a study at any point in time. Individuals withdraw from studies for many reasons, which may or may not relate to the study itself. Individuals may move or experience other personal issues that require withdrawal. Others may withdraw because they are no longer interested in the study, find the time commitment too great, or feel disappointed in the intervention. When people withdraw from a study, it can affect the equalization of groups that was achieved at the outset. When substantial numbers of participants withdraw from a study, group differences can emerge that confound the results of the study.

When attrition occurs, it is important to identify any characteristics of the individuals who dropped out of the study that might make them different from the individuals who remained in the study. Perhaps the individuals who dropped out were experiencing a more severe condition, in which case you would not know if the intervention was effective for that group of individuals. Attrition may also result in an uneven number of participants in the groups.

Protections Against Attrition/Mortality Threats

Depending on the length of the study and access to participants, a researcher may be able to recruit additional participants to replace individuals who drop out and thus maintain the overall power of the study. In addition, strategies such as reminder phone calls and e-mails can be used to promote good attendance for an intervention or follow-through with therapy.

Characteristics of the individuals who withdraw should be compared with those of the individuals who remain. If differences exist, this factor should be identified as a limitation of the study. Statistical procedures can be used to account for attrition/mortality threats, such as using estimates for missing data, but this approach is less desirable than having actual participant scores. An "intent to treat" analysis can be used in which the data of individuals who did not receive the intervention are still included in the analysis. This analysis is similar to real-life practice, in which some individuals do not complete or follow through with all aspects of their treatment. It also maintains the integrity of the randomization process and baseline equality of groups.

Exercise 5-2

Detecting Potential Threats to Internal Validity in a Research Study (LO3 and LO4)

Analyze the following two study abstracts and determine which threats to internal validity (among the options provided) are likely to be present. Before looking at the answers, write down a rationale for why you do or do not think a particular threat may confound the interpretation of the results. In other words, would the threat suggest that something other than the intervention resulted in the improvement? You can find the answers at the end of the chapter.

STUDY #1

Hobler, A. D., Tsao, J. M., Katz, D. I., Dipiero, T. J., Hebl, C. L., Leonard, A., . . . Ellis, T. (2012). Effectiveness of an inpatient movement disorders program for patients with atypical parkinsonism. Parkinson's Disease (2012), *871-974. doi:10.1155/2012/871974 (Epub 2011 Nov 10).*

Abstract

This paper investigated the effectiveness of an inpatient movement disorders program for patients with atypical parkinsonism, who typically respond poorly to pharmacologic intervention and are challenging to rehabilitate as outpatients. Ninety-one patients with atypical parkinsonism participated in an inpatient movement disorders program. Patients received physical, occupational, and speech therapy for 3 hours/day, 5 to 7 days/week, and pharmacologic adjustments based on daily observation and data. Differences between admission and discharge scores were analyzed for the functional independence measure (FIM), timed up and go test (TUG), two-minute

walk test (TMW), Berg balance scale (BBS) and finger tapping test (FT), and all showed significant improvement on discharge ($P > .001$). Clinically significant improvements in total FIM score were evident in 74% of the patients. Results were similar for ten patients whose medications were not adjusted. Patients with atypical parkinsonism benefit from an inpatient interdisciplinary movement disorders program to improve functional status.

Consider this:

- Not included in this abstract is the length of treatment. Participants' length of stay varied from 1 to 6 weeks, with an average of 2.5 weeks. Also, the intervention leaders administered the assessments.
- A working knowledge of atypical parkinsonism symptoms, course, and treatment will be useful in identifying threats to validity. You can obtain more information at http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001762/
- If you would like more information about the study, you may want to use your library resources to obtain the full text of the article.

QUESTIONS

1. Based on your reading of the abstract and any additional resources, which of the following would you consider noteworthy threats to internal validity?

   A. Maturation
   B. History
   C. Testing

   Explain your answer.

STUDY #2

Frankel, F., Myatt, R., Sugar, C., Whitham, C., Gorospe, C. M., & Laugeson, E. (2010, July). A randomized controlled study of parent-assisted Children's Friendship Training with children having autism spectrum disorders. Journal of Autism and Developmental Disorders, 40*(7), 827-842. doi:10.1007/s10803-009-0932-z*

Abstract

This study evaluated Children's Friendship Training (CFT), a manualized parent-assisted intervention to improve social skills among second to fifth grade children with autism spectrum disorders. Comparison was made with a delayed treatment control group (DTC). Targeted skills included conversational skills, peer entry skills, developing friendship networks, good sportsmanship, good host behavior during play dates, and handling teasing. At posttesting, the CFT group was superior to the DTC group on parent measures of social skill and play date behavior, and child measures of popularity and loneliness. At 3-month follow-up, parent measures showed significant improvement from baseline. Post-hoc analysis indicated more than 87% of children receiving CFT showed reliable change on at least one measure at posttest and 66.7% after 3 months follow-up.

Consider this:

- Ten participants did not complete the intervention and therefore were not included in the follow-up data.
- The following table was included in the study, comparing the two groups at baseline.

Sample Characteristics for Children's Friendship Training (CFT) and Delayed Treatment Control (DTC) Conditions

Variable	Group		*p*
	CFT *M* (SD) *n* = 35	DTC *M* (SD) *n* = 33	
Age (months)	103.2 (15.2)	101.5 (15.0)	ns
Grade	3.2 (1.0)	3.4 (1.2)	ns
SES[a]	44.6 (10.6)	50.6 (11.8)	ns
Percent male	85.7	84.8	ns
Percent Caucasian	77.1	54.5	ns
WISC-III Verbal IQ	106.9 (19.1)	100.5 (15.7)	ns
ASSQ	22.4 (7.3)	22.0 (9.3)	ns
VABS[b]			
Communication	84.3 (20.5)	79.8 (15.3)	ns
Daily living	67.0 (18.2)	62.4 (15.7)	ns

Continued

Sample Characteristics for Children's Friendship Training (CFT) and Delayed Treatment Control (DTC) Conditions *(continued)*

Variable	Group		*p*
	CFT *M* (SD) *n* = 35	DTC *M* (SD) *n* = 33	
Socialization	66.3 (10.8)	66.1 (10.8)	ns
Composite	68.1 (16.4)	64.4 (11.0)	ns
# sessions attended	11.3 (0.8)	10.7 (1.9)	ns

[a] DTC *n* = 32
[b] CFT *n* = 34

If you would like more information about the study, you may want to use your library resources to obtain the full-text article.

QUESTIONS

2. Based on your reading of the abstract and any additional resources, which of the following would you consider noteworthy threats to internal validity?

   A. Maturation
   B. Selection
   C. Instrumentation
   D. Attrition

   Explain your answer:

External Validity

External validity is the extent to which the results of a study can be applied to other people and other situations. External validity speaks to the generalizability of a study. A study has more external validity when it reflects real-world practice. As an evidence-based practitioner, those studies that include conditions that are more similar to your practice will have more external validity, and the results can be applied with greater confidence.

Threats to External Validity

Threats to external validity occur when the situation or study participants are different from the real world or the clinical setting. As with internal validity, external validity is a continuum; a study may have good or bad external validity, but it will never be perfectly valid. As an evidence-based practitioner, it is important that you evaluate the characteristics of the people and situations in a study to determine how similar those characteristics are to your own practice. Table 5-3 summarizes threats to external validity and protections against those threats.

TABLE 5-3 Threats to External Validity and Their Protections

Threat	Confounding Factor Affecting Generalizability	Protection
Sampling error	• Sample does not represent the population.	• Use random assignment. • Use large samples. • Select participants from multiple sites. • Replicate the study with new samples.
Poor ecological validity	• Conditions of the study are very different from real-world practice (when the research is administered in a manner that closely mirrors real-life practice, the results will be more generalizable).	• Ensure researcher is sensitive to issues of real-world practice. • Replicate with effectiveness studies.

Sampling Error

A primary principle of quantitative research involves generalizing the results from a sample to the population. **Sampling error** is any difference that exists between the population and the sample. The exact nature of sampling error is not always known, because many characteristics of the population are unknown. However, among known characteristics it is possible to compare the sample with the population to identify similarities and differences. For example, boys are approximately five times more likely than girls to be diagnosed with autism (CDC, 2012), although even this is an estimate from a sample. In a study that intends to represent children with autism, a more representative sample would be one with a similar gender distribution.

Protections Against Sampling Error

Sampling methods influence external validity. Although the best method for obtaining a representative sample is to randomly select a large sample from the target population, this is not always possible. In **random sampling** every individual in the population has an equal chance of being selected. With a large random sample, you are likely to select a group of participants that is representative of the population. Unfortunately, true random sampling rarely happens in health-care research, because it is usually impractical to sample from an entire population. For example, in an intervention study of people with multiple sclerosis, the population would be all individuals with multiple sclerosis. A worldwide sampling and then administration of the intervention would be next to impossible.

True random sampling does occur in some research when the sample is smaller and accessible. For example, a study of members of the American Occupational Therapy Association could be obtained through random sampling of the membership list.

The most common method of sampling in health-care research is **convenience sampling**. In this method, participants are selected because they are easily available to the researcher. When conducting a study of individuals with a particular condition or disability, it is likely that the researcher will go to a treatment facility that provides services to those individuals. Then the researcher might ask for volunteers, or a clinician might approach each person who meets the study criteria when that person is admitted. The lack of randomness in the process presents a high potential for introducing bias or sampling error. When samples are selected from one school, one neighborhood, or one clinic, for example, they are more likely to have characteristics that are different from the population as a whole; depending on the setting, they may be poorer, older, or more symptomatic.

One method for reducing sampling error is by selecting a large sample from multiple settings. A larger sample is more likely to approximate the population. In addition, multiple settings can be selected to represent the heterogeneity of a population. For example, in considering the generalizability of a study of children with attention deficit hyperactivity disorder (ADHD), it is more likely that a sample would represent the population if children from both urban and suburban locations in different areas of the country were recruited, to better represent the racial and socioeconomic characteristics of the population.

In the results section of a study, it is important for the researcher to provide a detailed description of the study participants. Many journals require that gender, age, and race at a minimum be included. As an evidence-based practitioner, you can review this information to determine if the sample is representative of the population. However, more important to you is whether the sample in the study is similar to the clients you work with. When a study sample is similar to your clientele, you are more justified in generalizing the findings.

Ecological Validity Threats

Ecological validity refers to the environment in which a study takes place and how closely it represents the real world. The treatment or method by which a study is administered, the time during which a study takes place, and where a study takes place are all important considerations affecting the external validity or generalizability of a study. Sometimes the administrators of the intervention in a study are highly trained, more so than the typical practitioner. The study time period may last longer than the length of stay covered by most insurance companies, or the intervention may be more intense than standard practice. The study may take place in an inpatient setting, although most individuals with the condition are actually treated on an outpatient basis. Any differences from the study conditions and real-world practice represent threats to external validity. The generalizability of a particular study will be good in situations that are similar to those in the study and poor in those that are different.

Protections Against Ecological Validity Threats

Practitioners are more likely to apply research that is relevant and practical to real-world practice, and studies have greater ecological validity when they are sensitive to typical practice situations. For example, a researcher may ensure that the intervention takes place in the typical time frame during which clients receive therapy, or that the therapists providing the intervention are those who already work in a particular type of hospital.

As a practitioner, it is important to apply the results of a study cautiously and consider the similarity to your

own situation. A study is more generalizable to your practice and clients when the characteristics of the study are similar. For example, a study by Sutherland et al (2012) that examined exposure therapy for posttraumatic stress disorder (PTSD) in veterans will be more applicable to practice situations that involve treating veterans with PTSD. The results of the study will be less applicable and have less external validity for treating PTSD in women who have experienced sexual abuse. In another example, a well-designed study that involved 24 weeks of Tai Chi showed that it was effective in improving balance for individuals with Parkinson's disease (Tsang, 2013). However, if you are unable to see clients for a 24-week time period, this study is less relevant for your practice setting. Nevertheless, you may be able to use the results of this study to justify to your administrators and/or insurance companies why a longer length of stay is warranted.

Replication to Promote Generalizability

Replication, or reproducibility, is essential to the generalization of research and a primary principle of the scientific method. It is important in the generalization of both samples and situations. Study findings must be capable of being repeated to ensure generalizability and applicability of the results. If several studies yield similar findings about the efficacy of an intervention or a predictor of an outcome, clinicians can be more confident in those results.

Another consideration in replication is the researchers themselves. Even if there are several studies that support a particular approach, if all of those studies were conducted by the same researcher, there should be some concern that the results will not generalize to other situations. There may be reasons why one researcher is able to garner more positive findings than another. Perhaps that researcher and his or her team are exceptional therapists and it is their general competence as opposed to the actual intervention that makes the difference.

In addition, when findings are particularly surprising or remarkable, replication is important. These kinds of findings are interesting, and thus will have a higher likelihood of being published. However, replication will reveal whether the findings were a result of chance (i.e., a Type I error).

Replication is often a matter of degree. Studies rarely follow the exact procedures of a previous study to determine whether the same results are obtained. Typically variables are manipulated to extend the findings of previous research. A replication study may shorten an intervention period, utilize a different outcome measure, apply the approach to a different sample, or administer the intervention in a new setting. The ability of research to build upon previous work is part of the power of the scientific method.

Internal Versus External Validity

When designing a study, the researcher must find a balance between internal and external validity. Studies that are tightly controlled to maximize internal validity will have less external validity. For example, inclusion criteria that produce a homogeneous sample, a strict protocol for administering the intervention, expert intervention

Evidence in the Real World

How Replication Changed the Perception of Facilitated Communication

In the early 1990s, there was great interest in facilitated communication for individuals with autism. Much of this interest came from the work of Biklen and colleagues (1992), who got surprising and amazing results with this technique. In facilitated communication, the facilitator provides physical assistance to help a person with autism type out a message on a keyboard. The assumption is that facilitated communication overcomes neuromotor difficulties that interfere with the abilities of a person with autism to communicate.

In an uncontrolled study of 43 individuals with autism, Biklen and colleagues reported startling outcomes. Previously nonverbal individuals were writing grammatically correct sentences and paragraphs, and even poetry. Skepticism about these findings led other researchers to conduct controlled studies. A review by Green (1994) found that when the facilitator's influence was controlled, the technique was no longer useful. The review suggested that the facilitator's belief in the potential of facilitated communication and the client's untapped capabilities led him or her to unconsciously or unintentionally guide the communication process. Now organizations such as the American Speech and Hearing Association and the American Psychological Association assert that there is no evidence to support facilitated communication for individuals with autism.

leaders, and limited exposure to alternative treatments will yield results that can be interpreted in the context of a cause-and-effect relationship, yet do not reflect everyday practice. In contrast, studies that are conducted under real-world conditions are "messier"; that is, there are not as many controls in place, and real-world conditions introduce more alternative explanations of the outcome—greater external validity at the expense of internal validity.

When accumulating research evidence regarding a particular intervention, you may use a process that begins with studies with high internal validity and moves to studies with greater external validity. First, it is important to know if an intervention is effective under ideal conditions; that is, a highly controlled study with strong internal validity.

Once the efficacy of the intervention is established, future studies can examine the same intervention in more typical practice conditions. The difference between these studies can be referred to as efficacy versus effectiveness. An **efficacy study** is one that emphasizes internal validity and examines whether an intervention is effective under ideal conditions. With efficacy studies, you can be more confident that the intervention is what made the difference; however, the conditions of the study are likely to differ from real-world conditions.

In an **effectiveness study**, the study conditions are more reflective of real-world practice; however, the untidy nature of practice means that there could be more threats to internal validity in play. Studies about therapy practices will always have threats to validity. Researchers face significant challenges in designing a study and must find a balance that involves minimizing threats, being pragmatic, and operating ethically. **From the Evidence 5-2** provides an example of an effectiveness study that carries out a strength training intervention in existing community fitness facilities.

Exercise 5-3

Managing Threats to Validity in a Particular Research Study (LO3, LO4, LO5, and LO6)

Childhood obesity is a major public health risk, and many efforts have been made to address the problem. A researcher is interested in studying a new intervention designed to increase the amount and intensity of physical activity for children in primary grades 1 through 3. Two schools have agreed to participate in the study. One school is located in an urban setting with children from mostly low socioeconomic and racially diverse backgrounds. Another school is located in a suburban setting with children from mostly high socioeconomic and Caucasian backgrounds.

QUESTIONS

Consider the following issues and describe how the researcher might reduce threats to validity. The following situations address both internal and external validity issues.

1. *The schools in which the researcher plans to implement the study will not allow the researcher to randomly assign children to groups.* What is the threat to validity, and how can the researcher manage this threat?

2. *The researcher plans to increase interest in physical activity by including a climbing wall and other new but expensive equipment as part of the activity program.* What is the threat to validity, and how can the researcher manage this threat?

3. *To determine if the activity at school carries over to home, parents are asked to keep a log for one week of their child's participation in activity, which includes type of activity, time engaged, and level of intensity.* What is the threat to validity, and how can the researcher manage this threat?

From the Evidence 5-2

An Example of an Effectiveness Study

Minges, K. E., Cormick, G., Unglik, E., & Dunstan, D. W. (2011). Evaluation of a resistance training program for adults with or at risk of developing diabetes: An effectiveness study in a community setting. International Journal of Behavioral Nutrition and Physical Activity, 8, *50. doi:10.1186/1479-5868-8-50.*

Note A: The researchers were interested in taking an intervention with efficacy in a controlled condition and assessing its effectiveness in existing fitness facilities.

Note B: The large number of dropouts (there were 86 participants at 2 months, but only 32 at 6 months) is not unexpected in a fitness center. People often discontinue their fitness program.

BACKGROUND:
To examine the effects of a community-based resistance training program (Lift for Life®) on waist circumference and functional measures in adults with or at risk of developing type 2 diabetes.

METHODS:
Lift for Life is a research-to-practice initiative designed to disseminate an evidence-based resistance training program for adults with or at risk of developing type 2 diabetes to existing health and fitness facilities in the Australian community. A retrospective assessment was undertaken on 86 participants who had accessed the program within 4 active providers in Melbourne, Australia. The primary goal of this longitudinal study was to assess the effectiveness of a community-based resistance training program, thereby precluding a randomized, controlled study design. Waist circumference, lower body (chair sit-to-stand) and upper body (arm curl test) strength, and agility (timed up-and-go) measures were collected at baseline and repeated at 2 months (n = 86) and again at 6 months (n = 32).

RESULTS:
Relative to baseline, there was a significant decrease in mean waist circumference (-1.9 cm, 95% CI: -2.8 to -1.0) and the timed agility test (-0.8 sec, 95% CI: -1.0 to -0.6); and significant increases in lower body (number of repetitions: 2.2, 95% CI: 1.4-3.0) and upper body (number of repetitions: 3.8, 95% CI: 3.0-4.6) strength at the completion of 8 weeks. Significant differences remained at the 16-week assessment. Pooled time series regression analyses adjusted for age and sex in the 32 participants who had complete measures at baseline and 24-week follow-up revealed significant time effects for waist circumference and functional measures, with the greatest change from baseline observed at the 24-week assessment.

CONCLUSIONS:
These findings indicate that an evidence-based resistance training program administered in the community setting for those with or at risk of developing type 2 diabetes can lead to favorable health benefits, including reductions in central obesity and improved physical function.

Note C: Without a control group, you could be less certain that Lift for Life made the difference. Perhaps individuals attending the fitness center took advantage of other programs or were more likely to exercise outside the program. Because the assessors are not blind to group assignment, they may consciously or unconsciously show a bias in scoring individuals whom they hoped were improving. This example demonstrates how improving external validity can sometimes compromise internal validity.

FTE 5-2 Question *Using the Lift for Life study example, how did improving external validity compromise internal validity?*

Critical Thinking Questions

1. Why is a large sample generally more desirable than a small sample in research (give at least three reasons)?

2. Why is a randomized controlled trial considered the strongest single study design?

3. Why might random assignment to groups result in ethical concerns?

4. Although pretests are generally desirable, how can they potentially pose a threat to validity?

5. For each of the following three situations, how can the researcher manage threats to validity to determine whether the new intervention is effective?
   - Comparing a new intervention with a no-treatment control group
   - Comparing a new intervention with a treatment-as-usual control group
   - Comparing a new intervention with another evidence-based intervention

6. Explain the differences between random selection and random assignment. What aspects of validity are addressed by these research practices, and how?

7. Why is it difficult to design a study that is strong in both internal and external validity? How can you balance the two types of validity?

Answers

EXERCISE 5-1

1. There are two reasons why fishing threats exist: The researcher does not have a research hypothesis, and four outcomes are being studied.
2. The study would be stronger if the researcher had a prior hypothesis about which orthoses would be best for which outcomes. This could be based on existing research or the researcher's clinical experience. To address the fact that multiple outcomes are studied, the researcher should adjust the alpha level of the statistical analysis or use a statistic to control for multiple comparisons.
3. Ten people divided into three groups will result in a study with very low power.
4. The researcher will want to recruit additional participants and may need to use another clinic or conduct the study over a longer period of time. Power can also be increased by reducing the number of groups, so the researcher could compare two orthoses (although it would still be best to have more than 5 participants per group) or use a crossover design in which all of the participants try all of the orthotics.

EXERCISE 5-2

1. Study #1
   A. Maturation—No. Although there is no control group and the study goes on for several weeks, consider the normal course of the disorder in determining whether or not maturation is a threat to validity. The normal course of Parkinson's disease is progressively deteriorating, so you would not expect improvement without treatment.

B. History—Yes. The conclusion of the study is that an interdisciplinary movement disorders program was effective in improving movement problems for people with Parkinson's disease. The medication adjustments could be a history threat: 81 of the 91 participants received a medication adjustment during the intervention. Although participants without medication adjustments had similar improvements, which provides some support for the therapy, the design of this study makes it difficult to determine whether it was the therapy or the medications (or both) that made the difference.

C. Testing—Yes. There are a few issues with testing. The Timed Up and Go Test uses time as an outcome, and the two-minute walk test is measured in terms of distance covered. With these objective outcomes, you would not be as concerned about biased assessments. However, the FIM and Berg Balance Scale do involve judgment on the part of the therapist, and a therapist who has been involved in the intervention and wants to see improvement may tend to rate the participants, albeit unintentionally, higher than an unbiased rater would. In addition, medication effectiveness in Parkinson's disease varies across the day, so the time of day at which assessments were administered could affect the outcome.

2. Study #2

A. Maturation—No, in this case there is a no-treatment control group with random assignment. If there was an improvement in the control group, the intervention's improvement was greater than the control's improvement and would suggest that the intervention group improved over and above any typical development.

B. Selection—No, random assignment helps to promote equal group assignment. The table provides additional support that the two groups were comparable at the outset.

C. Instrumentation—Yes. The self, parent, and teacher reports are problematic. The children, parents, and teachers knew about the intervention and may have been biased toward providing a more positive report. The use of observational methods (i.e., observing the child in social situations) would enhance this aspect of the study.

D. Attrition—Yes. Ten children in the intervention group were unavailable at the follow-up testing period: eight of these children dropped out, and two were removed for behavioral reasons. It is possible that these 10 children were not responding as well to the intervention and that could be why they dropped out. The two who were removed were not benefiting. If these children were included in the findings, it is possible, even likely, that the results would be less positive. This should be taken into account when evaluating how effective the intervention is and for how many.

EXERCISE 5-3

1. The major concern with lack of randomization is that there will be selection threats to validity. To address this concern, it is important to use strategies that will reduce any differences that might occur between the groups. In school-based research, it is common for one classroom to receive an intervention while the other classroom does not. You would not want to make one school an intervention setting and one school a control setting, because the distinct differences in the schools might account for differences you find in the intervention. Instead, you could randomly assign classrooms at each school to receive or not receive the intervention. You might address ethical concerns for the control group not receiving the intervention by using a wait-list control design (i.e., you will eventually provide the intervention to the control group). A drawback to this approach is that there is the potential for greater experimenter and participant bias. Blinding of the testers and reducing exposure of the students and teachers, particularly during physical activity, would help address these concerns. It would also be useful to provide the control group with equal attention to something new without introducing additional physical activity. For example, you might have the control group participate in board games.
2. The inclusion of expensive equipment makes it less likely that other schools will be able to implement this intervention, thereby making it less generalizable. The researcher should consider redesigning the intervention to use equipment that is typically available in most school settings; however, in doing so the researcher may lose the novelty or excitement that would be created by the new equipment.
3. Asking parents to keep a log introduces instrumentation threats. Maintaining a log for a week is asking a great deal of the parents, and it is unlikely that you will receive complete data. The researcher could use more objective means, such as an accelerometer that the children wear to record the time spent engaged in activity. Another method that is less burdensome to the parent is time sampling. In time sampling, usually at random intervals, a timer indicates that a log entry should be made. Using this method the parent only has to respond to the timer and not keep records at all times. Both of these methods still present concerns. For example, the parents may forget to put the accelerometer on, the child may lose the accelerometer, or the parent still may not respond to a time-sampling approach.

All of these examples speak to the challenges of designing a study. It is virtually impossible to design a perfect

study with no threats to validity. Researchers typically weigh their options and make choices given the particular research question, the ethical concerns presented, and pragmatic issues.

FROM THE EVIDENCE 5-1

1. Yes, the *p* value for all of the comparisons is > 0.05, indicating there is no statistically significant difference between the groups at baseline.
2. If the groups start out at different levels of back pain, this could affect/confound the results of the study. For example, if the control group had less pain and the intervention group had more pain at baseline, even without the intervention, maturation may result in the intervention group having a greater recovery, because there may be more room for improvement in the intervention group. The control group may not be able to improve much because they already are not experiencing a great deal of pain. In this example from the evidence, it is a good thing that the groups are equivalent on the outcome measure of back pain as well as other demographic variables.

FROM THE EVIDENCE 5-2

Without a control group, you could be less certain that Lift for Life made the difference. Perhaps individuals attending the fitness center took advantage of other programs or were more likely to exercise outside of the program. Also, you would expect less precision among the assessors, who would not be blind and may vary from site to site.

REFERENCES

Biklen, D., Morton, M. W., Gold, D., Berrigan, C., & Swaminathins, S. (1992). Facilitated communication: Implications for individuals with autism. *Topics in Language Disorders*, *12*(4), 1–28.

Centers for Disease Control and Prevention (CDC). (2012). Prevalence of autism spectrum disorders: Autism and developmental disability monitoring network, 14 sites, US, 2008. Retrieved from http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6103a1.htm?s_cid=ss6103a1_w

del Pozo-Cruz, B., Parraca, J. A., del Pozo-Cruz, J., Adsuar, J. C., Hill, J., & Gusi, N. (2012). An occupational, internet-based intervention to prevent chronicity in subacute lower back pain: A randomized controlled trial. *Journal of Rehabilitation Medicine*, *44*, 581–587.

Frankel, F., Myatt, R., Sugar, C., Whitham, C., Gorospe, C. M., & Laugeson, E. (2010, July). A randomized controlled study of parent-assisted Children's Friendship Training with children having autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*(7), 827–842. doi:10.1007/s10803-009-0932-z

Godi, M., Franchignoni, F., Caligari, M., Giordano, A., Turcato, A. M., & Nardone, A. (2013). Comparison of reliability, validity, and responsiveness of the Mini-BESTest and Berg Balance Scale in patients with balance disorders. *Physical Therapy*, *93*, 158–167.

Green, G. (1994). The facilitator's influence: The quality of the evidence. In H. C. Shane (Ed.), *Facilitated communication: The clinical and social phenomenon* (pp. 157–226). San Diego, CA: Singular.

Killen, J. D., Fortmann, S. P., Newman, B., &Varady, A. (1990). Evaluation of a treatment approach combining nicotine gum with self-guided behavioral treatments for smoking relapse prevention. *Journal of Consulting and Clinical Psychology*, *58*, 85–92.

Mayo, E. (1949). *Hawthorne and the Western Electric Company: The social problems of an industrial civilisation*. London, UK: Routledge.

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., & Fisher, P. (2007, July 3). The Hawthorne effect: A randomised, controlled trial. *BMC Medical Research and Methodology*,7, 30.

Minges, K. E., Cormick, G., Unglik, E., & Dunstan, D. W. (2011, May 25). Evaluation of a resistance training program for adults with or at risk of developing diabetes: An effectiveness study in a community setting. *International Journal of Behavioral Nutrition and Physical Activity*, *8*, 50. doi:10.1186/1479-5868-8-50

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York, NY: Holt, Reinhart & Winston.

Sabbag, S., Twamley, E. W., Vella, L., Heaton, R. K., Patterson, T. L., & Harvey, P. D. (2012). Predictors of the accuracy of self assessment of everyday functioning in people with schizophrenia. *Schizophrenia Research*, *137*, 190–195.

Sammons Preston. (n.d.). *Jamar hand dynamometer owner's manual.* Retrieved from https://content.pattersonmedical.com/PDF/spr/Product/288115.pdf

Sutherland, R. J., Mott, J. M., Lanier, S. H., Williams, W., Ready, D. J., & Teng, E. J. (2012). A pilot study of a 12-week model of group-based exposure therapy for veterans with PTSD. *Journal of Trauma and Stress*, *25*(2), 150–156.

Tsang, W. W. (2013). Tai Chi training is effective in reducing balance impairments and falls in patients with Parkinson's disease. *Journal of Physiotherapy*, *59*, 55.