"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
KEY TERMS
INTRODUCTION
VALIDITY
STATISTICAL CONCLUSION VALIDITY
Threats to Statistical Conclusion Validity
Fishing
Low Power
INTERNAL VALIDITY
Threats to Internal Validity
Assignment and Selection Threats
Maturation Threats
History Threats
Regression to the Mean Threats
Testing Threats
Instrumentation Threats
Experimenter and Participant Bias Threats
Attrition/Mortality Threats
EXTERNAL VALIDITY
Threats to External Validity
Sampling Error
Ecological Validity Threats
INTERNAL VERSUS EXTERNAL VALIDITY
CRITICAL THINKING QUESTIONS
ANSWERS
REFERENCES

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.

5

“世界不會停下來思考——它永遠不會，這不是它的方式;它的方法是從單個樣本中推廣出來。
——馬克吐溫

有效性

是什麼讓研究變得強大？

本章大綱

學習成果關鍵術語

介紹

有效性

統計結論有效性 對統計結論有效性的威脅

釣魚低功率

INTERNAL VALIDITY 對內部有效性的威脅

分配和選擇威脅

成熟威脅 歷史威脅

回歸到均值威脅 測試威脅

檢測威脅

實驗者和參與者偏見威脅

流失/死亡威脅

外部有效性 對外部有效性的威脅

抽樣誤差 生態有效性 威脅

內部有效性與外部有效性

批判性思維問題

答案 參考文獻

學習成果

1. 檢測已發表研究中對統計結論有效性的潛在威脅。
2. 在給定的研究中，確定研究人員是否充分管理了對統計結論有效性的潛在威脅。
3. 檢測已發表研究中對內部有效性的潛在威脅。
4. 在給定的研究中，確定研究人員是否充分管理了對內部效度的潛在威脅。
5. 檢測已發表研究中對外部有效性的潛在威脅。
6. 在給定的研究中，確定研究人員是否充分管理了對外部效度的潛在威脅。

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 whys 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.

關鍵術語

替代治療威脅

作業威脅

磨損

Bonferroni 校正

補償

士氣低落

治療的補償均衡

便利採樣

係數

生態效能

有效性研究

功效研究

實驗者偏差

外部有效性

#shing

霍桑效應

歷史威脅

檢測威脅

內部有效性

匹配

成熟威脅

死亡率

順序效果

參與者偏倚

權力

練習效果

皮格馬利翁效應

隨機分配

隨機抽樣

回歸均值

複製

回復率

羅森塔爾效應

採樣誤差

選擇威脅

統計結論效度

測試效果

有效性

介紹

本章的目的是確保您不會按照馬克吐溫在開場白中建議的那樣，而是從單個研究樣本中進行概括。您將學習如何停下來思考來自單個樣本和多個樣本的數據，目標是在循證實踐中使用這些資訊。在評估證據的強度時，從業者傾向於依賴某些公理，例如「隨機對照試驗是最強的設計」和「大樣本量提供更可靠的結果」。儘管在許多情況下都是如此，但也可能存在例外情況。要成為研究的批判性消費者，瞭解這些斷言背後的原因至關重要。為什麼需要大樣本量？為什麼隨機對照試驗中固有的保護措施？即使進行大樣本、隨機、對照試驗，其他因素也可能損害研究的有效性。

本章解釋了有效性的概念，描述了對有效性的不同威脅，並確定了針對這些威脅的可能解決方案。這些資訊將提高您批判性評價研究的能力。如果您很好地掌握了對有效性的可能威脅以及管理這些威脅的方式，您將能夠評估證據的強度並成為一名基於證據的專業人士。

有效性

在考慮研究的有效性時，請考慮真實性、健全性和準確性等術語。有效性是指導研究的設計、實施和解釋的研究理想。當合理的方法讓消費者對研究結果充滿信心時，研究的有效性就會增強。當得出的結論基於對統計數據的準確解釋而不是與其他解釋混淆時，研究的有效性就得到了支援。如果從研究中得出的推論可信並反映事實，那麼它們將具有更大的有效性。本章描述了三種類型的研究效度：（1）統計結論效度，（2）內部效度，以及（3）外部效度。第 6 章討論了與研究人員使用的評估有關的不同類型的效度。

統計 結論 有效性

統計結論效度是指從研究的統計分析中得出的結論的準確性。回想一下，在大多數推論統計中，計算的是 p 值;通常，如果 P 值< 0.05，則結論具有統計顯著性（即存在統計學顯著性差異或存在統計學顯著性關係）。作為一名循證從業者，您可能有理由質疑研究文章中提出的研究人員的結論。

對統計結論有效性的威脅

在第 3 章中，根據 I 類和 II 類錯誤描述了錯誤的統計結論。作為循證從業者，您可以通過提高對導致錯誤的研究實踐的認識來識別潛在的錯誤。本章介紹了統計結論有效性的具體威脅、它們與錯誤類型的關係以及研究人員用來保護研究免受這些威脅的方法。表 5-1 概述了統計結論有效性的威脅、干擾統計結論的混雜因素以及防範這些威脅的方法。

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,

表 5-1 對統計結論有效性的威脅及其保護

威脅	錯誤類型	干擾統計結論的混雜因素	保護
釣魚類型 I	• 研究人員在數據中搜索超出初始假設的有趣發現。 • 結論可能是偶然的。		• 使用針對多個分析進行調整的統計方法。• 進行第二項研究，與不同的參與者一起檢驗新假設。
低功耗	II 型	• 存在差異或關係，但沒有足夠的統計能力來檢測它。	• 提高 Alpha 等級。 • 確保干預得到充分實施以獲得最佳效果大小。 • 增加樣本量。

釣魚

釣魚是一種委婉的說法，指的是尋找研究人員最初不打算探索的發現。理想情況下，當研究人員進行研究時，會在收集數據之前提出假設。收集數據后，將應用統計分析來檢驗假設。然而，研究人員通常會以有時被稱為「釣魚探險」或「挖掘數據」的方式探索現有數據。換句話說，研究人員讓數據引導我們得出有趣的發現。儘管有正當理由深入研究數據，但釣魚的風險在於研究人員會看到有趣的差異或關係，這些差異或關係可能不是真的，而只是由於偶然。換句話說，研究人員通過發現不存在的差異而犯了 I 類錯誤。

通常，在研究人員搜尋結果時會進行許多分析。之前，您瞭解到，當alpha設置為0.05時，研究人員願意承擔5%的風險，即差異或關係不是真的，而是由於偶然性。但是，這僅適用於單個分析。每次執行另一次分析時，發現是由於偶然性的可能性更大。研究人員經常探索他們的數據以尋找意想不到的發現，這可能會導致重要的發現。但是，應該採取保護措施，以免偶然發現產生誤導。

一個直截了當的研究人員可能會承認這項探索，如果這是一項強有力的研究，將描述如何解決對 I 類錯誤的威脅。

研究人員防範捕撈威脅的一種方法是使用考慮多重分析的統計程式。這樣的過程有很多，但從概念上講，最簡單的是 Bonferroni 校正。使用 Bonferroni 校正時，通過將 alpha 水準 0.05 除以比較次數來調整 alpha 水準。例如，如果進行了六次比較，則0.05/6 =0.0083，這意味著可接受的alpha率比最初的0.05低得多，也更保守。

另一種防止釣魚威脅的方法涉及進行另一項研究，以檢驗在探索數據時發現的新假設。例如，考慮一位研究人員測試了一種新的干預措施，發現初步分析並未顯示干預措施比對照組更有效。然而，經過更深入的分析，研究人員發現男性體驗到了顯著的好處，而女性則保持不變。可以進行一項新的研究來檢驗這一假設。如果第二項研究得出相同的結果，將有更有力的證據得出結論，即只有男性才能從干預中受益。

防範捕魚威脅

作為一名循證從業者，當研究結果沒有以研究假設的答案來呈現時，您可能會懷疑發生了一次釣魚探險

低功耗

功效是研究檢測差異或關係的能力。功效基於三個因素：樣本大小、效果大小和Alpha級別。樣本越大，研究越強大。當您有許多參與者時，更容易檢測到差異。同樣，如果你有很大的效果，你就會有更大的力量。如果干預對結果產生重大影響，

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?

與干預僅產生微小差異相比，更容易發現這種差異。

回想一下第 3 章，當沒有發現差異，但實際上存在差異時，就會發生 II 類錯誤。發生這種情況是因為功效低，並且通常是樣本量小的結果。當您審查樣本量較小的研究，但未發現差異或關係時，低功效是對統計結論有效性的潛在威脅。然而，也有可能，即使樣本很大，研究人員也找不到差異或關係。

抵禦低功耗威脅

可以通過更改 Alpha 級別、效果大小或樣本大小來增加功效。在探索性分析中，研究人員可能會使用更高的 alpha 水準，例如 0.10 而不是 0.05;但是，這樣做時，研究人員犯 I 類錯誤的機會更大。改變效應大小更加困難，但研究人員需要確保一切都準備就緒，以測試干預措施是否有效（例如，受過培訓的個人進行干預，使用促進依從性的策略等）。

增加檢驗功效的最簡單方法是增加樣本量。然而，使用大量樣本進行研究在時間和資源方面都可能很昂貴。研究人員通常會進行功效分析，以確定在給定設定的alpha水平和估計的效應大小的情況下，可以檢測到效應的最小樣本。

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

在考慮其他分數時，分數 26 是一個異常值。這個樣本的平均分數為 6 分，為 8。當您查看每個參與者時，8 分比大多數參與者獲得的分數要高得多。

單個異常值會錯誤地表示整個組。

現在考慮 40 個參與者的樣本：

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

異常值對整個組的影響較弱，並且該樣本的平均值更符合典型分數。

大樣本的另一個好處是，樣本越大，樣本代表總體的可能性就越大。這一事實與調查研究特別相關。不僅大型 sample 代表總體，但它表明，當受邀完成調查時，可能會有更多的人做出回應。回應率是響應參與調查或其他研究工作的請求的人數。回應越大，結果越準確。在調查研究的情況下，回復率是通過將完成的調查數量除以管理的調查數量來確定的。例如，如果發出了 200 份調查，其中 150 人完成並返回了這些調查，則回復率將為 150/200 = 75%。

選擇不參加研究的個人可能會出於特定原因退出宣告研究，這樣做會使結果產生偏差。例如，如果您正在為您的治療計劃進行滿意度調查，但只有 25% 的客戶做出回應，那麼做出回應的人可能是非常不滿意或非常滿意，因此更有動力表達他們的意見。

練習 5-1

識別對統計結論效度的威脅（LO1 和 LO2）

閱讀以下場景並確定哪些做法對統計結論有效性構成潛在威脅。建議控制這些威脅的方法。

一位擔任治療師的新研究人員希望收集數據，以檢查三種不同矯形器對特定手部疾病的療效。研究人員計劃從她的診所招募客戶，並預計大約有10人患有感興趣的手部疾病。將測量以下結果：疼痛、運動範圍、精細運動控制和功能。研究人員對哪種矯形器會提供更好的結果沒有期望。

問題

1. 為什麼存在捕魚威脅？

2. 研究人員如何應對 fishing 威脅？

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

3. 為什麼會因低功耗而存在威脅？

4. 研究人員如何解決低功耗威脅？

內部 有效性

在評估研究的證據時，有必要考慮內部有效性以及它如何影響研究結果。當從結果中得出的結論準確和真實時，研究具有內部有效性。有效性不是一個非此即彼的情況，而是一個程度的問題。例如，一項檢查社交故事對自閉症兒童有效性的研究得出結論，干預組兒童的社交技能比對照組兒童有更大的進步。如果這項研究在內部是有效的，這將意味著確實是社交故事干預提高了社交技能。

然而，研究結果總是有可能有其他解釋。也許這種差異是由於偶然。或者可能是孩子們受到的關注造成了差異，而不是干預本身。也許進行結果評估的個體有偏倚，並且傾向於給干預組中的個體更高的分數。儘管您永遠無法確定研究的結果是否完全準確，但研究的某些特徵可以大大提高您對其準確性和有效性的信心。

對內部有效性的威脅

在檢查內部效度時，問問自己，「這些研究結果有沒有其他解釋？其他解釋通常被稱為對內部有效性的「威脅」。本章的這一部分描述了對內部有效性的常見威脅，描述了避免或最小化這些威脅的保護措施或解決方案，並確定了這些威脅最有可能發生的研究情況類型。表 5-2 總結了對內部有效性及其保護的威脅。

分配和選擇威脅

如果在分配或選擇過程中存在偏差，則可能會對內部有效性構成威脅

表 5-2 對內部有效性的威脅及其保護

威脅	影響結果的混雜因素/替代解釋	保護
成熟	• 由於發育或癒合，參與者會隨著時間的推移而發生變化。	• 使用對照組。 • 通過隨機分配或參與者匹配確保基線等效性。
分配/選擇	• 群體在某些重要特徵上並不相等。	• 隨機分配 • 參與者匹配 • 協方差等統計程式
歷史	• 事件發生在前測和后測之間。	• 使用對照組。 • 確保前測和后測之間的時間較短。• 確保防止接觸替代療法。

繼續

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

表 5-2 對內部有效性的威脅及其保護措施（續）

威脅	影響結果的混雜因素/替代解釋	保護
回歸到均值	• 極端分數會隨著重複測試而變化並趨向平均值。	• 使用對照組。 • 排除異常值。 • 取多次測量的平均值。
測試/練習/順序效果	• 由於暴露或測試體驗的其他某些特徵，度量的性能會發生變化。	• 使用對照組。 • 使用具有良好測試/重測可靠性的度量 • 使用替代形式的度量。 • 平衡措施的順序。 • 如果預計會感到疲勞，請休息一下。
儀錶	• 無效或不可靠的措施、測試人員錯誤或儀器狀況不佳會導致結果不準確。	• 使用具有良好可靠性和有效性的措施。• 使用對變化敏感的措施。 • 培訓測試人員。 • 維護儀器。 • 使測試儀失明。
參與者和實驗者偏差威脅		
羅森塔爾/皮格馬利翁效應	• 干預領導者希望參與者有所改善，或者參與者期望有所改善。• 當涉及對照組時，預期實驗組的表現會比對照組好。	• 對干預領導者和/或參與者實施盲法。• 限制干預和控制領導者與參與者之間的接觸。 • 建立並遵循干預方案。
補償均衡治療	• 干預領導者的行為鼓勵對照組的參與者提高到與干預組持平，或者對照組參與者有動力與干預組競爭。	• 盲法干預領導者和/或參與者。• 限制干預和控制領導者與參與者之間的接觸。
補償士氣低落	• 干預領導者的行為使對照組感到氣餒，或者參與者因為在對照組中而氣餒。	• 盲法干預領導者和/或參與者。• 限制干預和控制領導者與參與者之間的接觸。

繼續

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

表 5-2 對內部有效性的威脅及其保護措施（續）

威脅	影響結果的混雜因素/替代解釋	保護
霍桑影響	• 參與者因為參與研究而受到關注而有所提高。	• 盲法干預領導者和/或參與者。• 確保對干預組和對照組給予同等關注。
損耗/死亡率	• 退出的參與者會影響群體的均衡或參與者的其他特徵。	• 採用策略來鼓勵出勤或參與。• 對缺失數據使用統計估計。 • 執行意向治療分析。

群組的參與者。如果各組之間的基線特徵存在差異，則這種差異可能會影響研究的結果。

人口統計學特徵、疾病/狀況問題、藥物和結果測量的基線分數是在檢查分配和選擇威脅時應考慮的變數示例，因為這些特徵可以解釋結果的差異。分配威脅表示偏差發生在分配組時，而選擇威脅表示偏差發生在選擇過程中——無論是選擇參與者還是選擇地點。例如，Killen、Fortmann、Newman 和 Varady（1990）發現男性對特定戒煙計劃的反應優於女性。如果在基線時干預組男性更多，對照組女性更多，則研究將偏向於發現比兩組性別分佈相等時更多的積極結果。在另一個例子中，腦癱的夾板研究可能會發現一組接受的抗痙攣藥物比另一組更多。然後，藥物可以解釋結果中的部分或全部差異。

基線時干預組和對照組的比較應在研究結果部分的初始部分提供，以便讀者可以確定組間是否存在差異。通常會提供一個表格，概述各組在重要人口統計學特徵以及結果變數的基線分數方面的比較。這種比較的一個示例顯示在From the Evidence 5-1 中。該表來自一項干預研究，該研究檢查了教育和鍛煉計劃對減少腰痛慢性化的有效性該表比較了基線時的干預組和對照組（del Pozo-Cruz 等人，2012 年）。

防範作業威脅

隨機分配是研究人員用來抵禦分配威脅的主要保護方法。在隨機分配到組時，每個研究參與者都有平等的機會被分配到可用於干預或控制的組。有時研究人員不願意將隨機分配到無干預對照組，因為拒絕對一組進行治療可能被認為是不道德的。有時，通過使用候補名單控制組來管理此問題。對照組最終接受干預，但要等到干預組完成治療后才能接受干預。

隨機分配並不能確保均勻分佈，對於小樣本尤其如此，因為少數個體的極端情況會極大地影響組結果。但是，它特別適用於較大的樣本，因為您更有可能形成等效組。因此，在評估證據時，請檢查研究結果部分提供的組比較。

有時，研究人員使用諸如匹配研究參與者之類的策略來確保在特別重要的特徵上平均分配。例如，如果研究人員知道結果可能受到教育水準、癥狀嚴重程度或藥物等特徵的影響，則會根據感興趣的變數確定並匹配潛在參與者，其中一名隨機分配到干預組，一名隨機分配到對照組。

另一個可用於最小化分配威脅的過程是組的統計均衡。如果

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

來自證據 5-1

表格 比較研究組

del Pozo-Cruz， B.， Parraca， J. A.， del Pozo-Cruz， J.， Adsuar， JC， Hill， J.， & Gusi， N.（2012）.一種基於互聯網的職業干預，用於預防亞急性腰痛的慢性病：一項隨機對照試驗。康復醫學雜誌，44（7），581-587。

doi：10.2340/16501977-0988.

表 I. 研究參與者的基線特徵（n = 90）

群	控制組（n = 44）	干預組（n = 46）	p
	平均值（SD）	平均值（SD）	
年齡（歲）	45.50 (7.02)	46.83 (9.13)	0.44
性別（%）			
Male	11.4	15.2	
女性	88.6	84.8	0.59
吸煙者，是/否，% 羅蘭莫裡斯問卷評分，TTO 分，SBST 總分，SBST 心理評分，分數	50/50	56.5/43.5	0.53
	11.65 (2.14)	12.28 (2.63)	0.22
	0.78 (0.08)	0.75 (0.11)	0.23
	4.38 (1.67)	4.36 (1.28)	0.95
	2.36 (1.03)	2.28 (0.98)	0.70

來自獨立測量或 2 檢驗的 t 檢驗的 p 值。
TTO：時間權衡;SBT： STarT Back Tool（STarT 後背工具）;SD：標準差。

注 A：SBST 是該研究的結果測量。

注 B：所有比較的 p 值均高於 0.05，表明兩組在基線時具有可比性（即，沒有統計學上的顯著差異）。SBST 總分尤其如此。

FTE 5“1 問題 1 這兩個群體在所有關鍵特徵（人口統計變數和結果變數）上是否相同？你怎麼知道？

一組比另一組年長，年齡可以在統計分析中協變，因此不會影響結果。如果在閱讀結果部分的初始部分時，您發現各組在一個或多個重要特徵上不相等（有時會發生，即使隨機分配），請檢查研究人員是否通過協變該變數來處理此問題，或者至少在討論的局限性部分承認差異。

成熟威脅

成熟是涉及醫療保健從業人員的干預研究中的潛在威脅。成熟是指研究參與者隨著時間的推移而發生的變化。

在醫療保健研究中，兩種主要類型的成熟威脅特別常見：（1）作為自然生長過程的一部分發生的變化，這與兒童研究特別相關;（2）自然癒合過程發生的變化，這與與預期恢復的疾病和病症相關的研究尤其相關。換句話說，如果放任不管，研究參與者是否有可能自行發生變化？當前測和后測之間的時間延長時，例如在縱向研究或長期隨訪研究中，成熟度是最受關注的。

為了說明成熟威脅，請考慮一項研究，該研究檢查了對語言發育遲緩兒童的干預措施。研究發現，語言的改善來自

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?*

前測到後測;然而，如果沒有充分的保護免受其他影響，很難確定是干預導致了改善，還是由於語言發育的變化而發生了變化。如果研究持續很長一段時間，例如整個學年，成熟將是一個更大的問題。同樣，一項檢查髖關節置換術后個體活動能力變化的干預研究需要考慮成熟威脅，因為個體可以在沒有治療的情況下體驗到活動能力的改善。

無論自然變化是積極的還是消極的，成熟都在起作用。當條件導致自然衰退時，治療的目標通常是減少衰退發生的速度。例如，如果治療師正在對阿爾茨海默病患者進行認知干預，那麼確定衰退是否比在病程中自然發生的要輕，這將是具有挑戰性的。然而，可以設計具有適當保護措施的研究，以確定特定干預措施是否會減少功能下降的自然過程。

針對成熟威脅的保護

針對成熟威脅的主要保護措施是使用控制組。如果干預組比對照組改善得更多，則兩組之間的差異更有可能是由於干預造成的，即使兩組都隨著時間的推移而改善。干預組高於對照組的改善程度可能是由於干預。

另一種防止成熟威脅的措施是對照組和干預組在基線時的結果評分相似。這使您可以更確信這些組從相似的位置開始，並使對從 pretest 到 posttest 的更改的解釋更加簡單。

隨機分配和參與者匹配是增加

組在baseline上將相等。（隨機分配和匹配在 分配和選擇威脅一節中有詳細介紹。在研究的結果部分，通常結果的第一份報告是干預組和對照組的比較;這包括對可能影響結果的興趣和人口統計變數的預測試分數。最後，您可以更加確定，當前測和后測之間的時間很短，並且在沒有干預的情況下不太可能發生變化時，成熟不是一個因素。

歷史威脅

歷史威脅涉及由於前測和後測之間發生的事件而導致的結果或因變數的變化，例如參與者接受意外的治療或暴露於影響研究結果的活動。在這種情況下，該威脅也可以稱為替代治療威脅。例如，預防跌倒計劃的參與者可能會開始參加一個新的老年中心，該中心提供強調力量和平衡的鍛煉課程。

事實上，任何可能影響因變數的外部事件都是潛在的威脅。課堂上使用創新方法的新教師、參與引起對特定健康實踐的關注的健康調查或在參與者附近開設新的健身中心都可能對內部效度構成威脅。

病史也會對結果產生負面影響。

暴風雪可能會影響出勤率，或者在感恩節和耶誕節假期前後安排減肥計劃可能會干擾預期的結果並對內部有效性構成威脅。

針對歷史威脅的保護

許多用於防止成熟效應的相同策略可以避免歷史威脅。使用對照組可提供保護，只要兩個組對歷史

來自證據 5-1（續）

FTE 5“1 問題 2 使用此示例，為什麼干預組和對照組的參與者在 STarT Back Tool 的基線得分相似很重要？基線等效性如何防止成熟威脅？

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

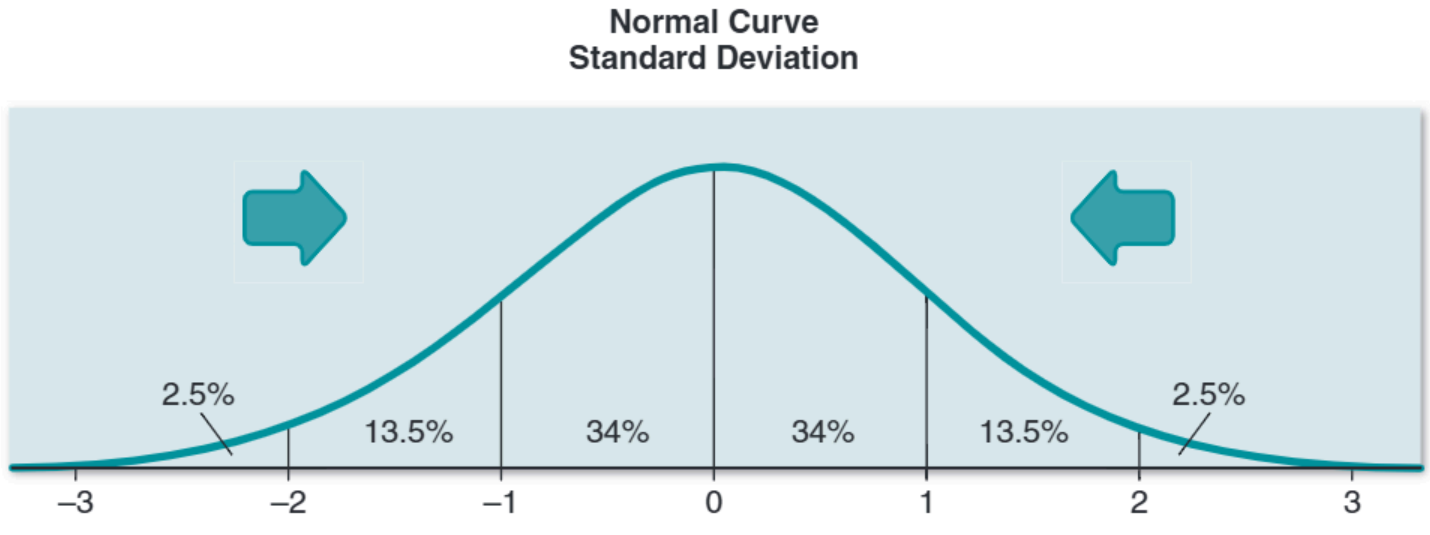

FIGURE 5-1 Normal curve, showing standard deviations and propensity for extreme scores to regress toward the mean. The shape of the curve suggests that individuals who score toward the ends are more likely to move toward the middle on a second testing.

事件。同樣，當前測和後測之間的時間較短時，歷史就會被簡化為一種威脅。研究人員還可以採取保護措施來減少對替代療法的暴露，例如要求參與者在參與研究時避免替代鍛煉計劃或藥物治療。研究人員可以包括調查問卷或觀察結果，以幫助確定是否發生了可能影響結果的事件。

回歸到均值威脅

回歸均值是指在進行第二次測量時，極端分數可能會向平均值移動的現象;極高的分數會變得更低，極低的分數會變得更高。第二次參加考試時，您總是有可能（甚至很可能）不會獲得完全相同的分數。這種現象在最初得分處於分佈的極端值的個體中尤其可預測。在分佈的末端，第二個測試分數變得更極端的可能性更小;相反，極端分數往往會回歸到平均值。“體育畫報詛咒”就是一個很好的例子。經常觀察到，在某人出現在《體育畫報》上后，這個人的表現會下降。回歸均值可以解釋這一觀察結果，因為當單個運動員處於最佳表現並且優於該運動中的大多數（如果不是所有其他）運動員時，他或她可能會成為特色。因此，後續性能可能會趨向於平均水準，而不是提高。

在醫療保健研究中，由於他們的病情，研究參與者通常以極高的分數開始。因此，當涉及極端分數時，應將回歸平均值視為潛在威脅。圖 5-1 描繪了正態曲線，並說明瞭極值分數回歸到平均值的傾向;連續體兩端的極值分數向中間移動。

再一次，如果治療組的表現優於對照組，則組間差異很可能是由於干預造成的。對照組的重要性應該越來越明顯;控制組很有價值，因為它們解決了有效性的多個威脅。

另一種選擇是研究人員從研究中排除異常值，儘管當大量參與者可能被歸類為異常值時，這種策略是不可行的。當需要小樣本時，可以通過採取多種預測試措施並使用平均值來減少基線極值分數帶來的威脅。例如，準確測量腰圍可能具有挑戰性，因此在測試過程中，可能會採取三個測量值，然後取平均值。

測試威脅

當測試性能的變化是測試體驗本身的結果時，就會發生作為內部有效性威脅的測試。當早期體驗以某種方式影響後續測試體驗時，就會出現測試效果。這種情況可能以許多不同的方式發生。

測試經驗通常會使參與者對理想的結果敏感。例如，前測可能會詢問有關遵循家庭計劃的問題，因此個人對這種行為變得敏感並開始遵循該計劃（作為測試的結果，而不是干預）。在另一個例子中，計步器和其他設備通常用作身體活動的量度。佩戴計步器的簡單動作可以影響一個人走多遠，因為計步器的存在會激勵人走得更多，尤其是當參與者可以看到讀數時。在這種情況下，是計步器的佩戴而不是干預導致了變化。測試人員還可以通過行為來影響測試結果，例如提供提示以提高性能，例如，“更努力，你可以再做一些”。

練習效應是一種測試威脅，當暴露於前測可以使個人在後測中表現更好時發生。先前的接觸可能意味著測試更熟悉，參與者

防止回歸均值威脅

與歷史和成熟威脅類似，使用對照組可以防止回歸均值。

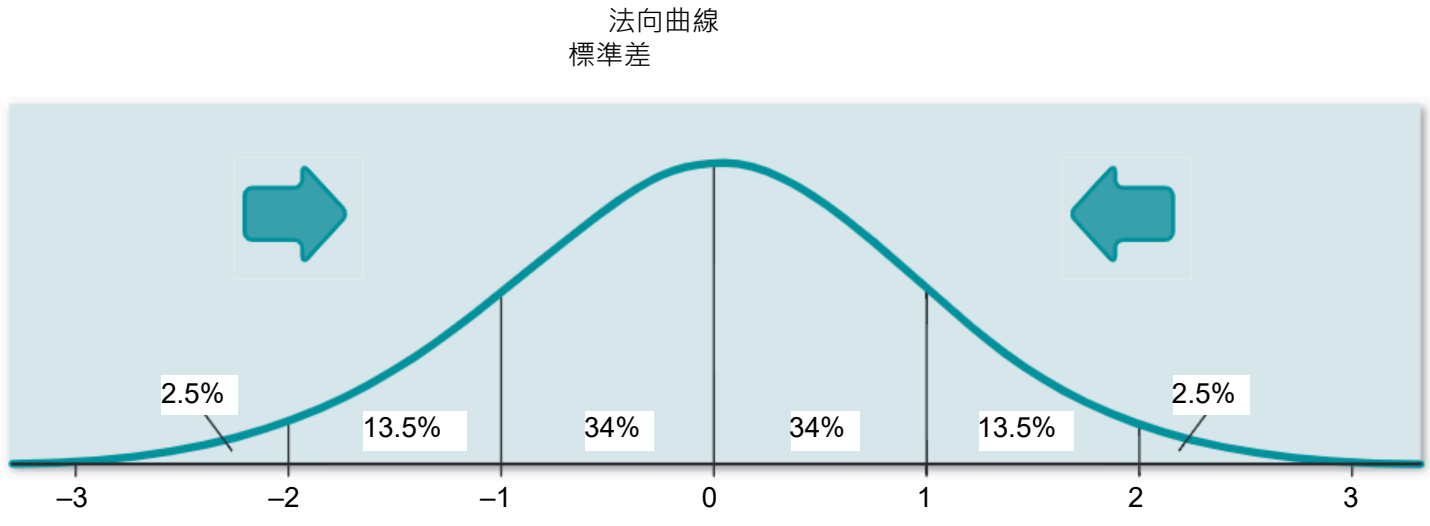

圖 5“1 正態曲線，顯示標準差和極值分數回歸平均值的傾向。曲線的形狀表明，在第二次測試中，得分偏向終點的個體更有可能向中間移動。

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

不那麼焦慮，參與者可以在後測中採用提高表現的策略。例如，在筆跡干預之前和之後接受筆跡測試的學生可能在第二次測試中表現更好，這僅僅是因為第一次測試的接觸和練習。

對於順序效應，性能會根據測試的顯示順序而變化。例如，在長時間的測試會話中，可能會因疲勞而導致性能下降。

防範測試威脅

具有很強的重測信度的度量是防範測試威脅的良好開端。標準化、腳本和測試人員的培訓也可以減少測試人員可能引入的偏差。時間採樣可用於防範測試威脅，而不是可以用作參與行為提示的日誌或日記。通過時間採樣，個人會收到一個提示，但提示是意外和隨機的，個人或工具會記錄該人在該時間點正在做什麼。對照組也有利於測試威脅。如果兩組都從暴露於測試情況中獲得一些益處，則對照組和干預組之間的差異仍然代表干預效果。

有些措施比其他措施更容易受到實踐影響。例如，評估記憶力的清單學習測試在第二次進行時不太可靠，因為參與者可能會記住第一次的單詞。相比之下，運動範圍測試不太適合實際效果。當可以學習度量時，通常會使用替代形式，例如單詞清單。在這種情況下，可以使用替代清單，以便使用一組不同的單詞來測試個人。

檢測威脅

當度量值本身或管理該度量值的個人不可靠或無效時，就會發生檢測威脅。檢測威脅是研究中的常見問題。如果 instrumentation 對其有效性構成威脅，則設計良好的研究將變得毫無用處。當使用機械或電子測量時，器械的品質、狀況和校準會影響結果。當儀器狀況不佳時，測量值可能不準確。例如，建議每年對 Jamar 測功機進行專業校準（Sammons Preston， n.d.）。

人為錯誤也可能在檢測威脅中發揮作用。例如，測試人員可能會向參與者提供不正確的指令，或者在對觀察性測量進行評分時做出糟糕的判斷。該測試本身對於研究來說可能是一個糟糕的選擇，例如當它不能準確測量預期結果時;這代表了測試有效性的問題，這是干預研究的一個非常重要的特徵。例如

許多治療研究使用自我報告措施來評估個體的功能表現。儘管這些措施易於管理，但它們可能無法提供準確的結果。Sabbag 等人（2012）發現，在評估精神分裂症患者的日常生活技能時，績效測量比自我報告更準確。

檢測威脅的另一個方面是度量的「天花板效應」或「下限效應」。。如果天花板效應在起作用，參與者一開始可能會有如此高的分數，以至於沒有改進的餘地。或者，測試可能非常困難（地板效應），以至於研究人員不太可能檢測到顯著變化。

防範檢測威脅

在選擇受試者之前，正確選擇措施是防止儀器威脅的重要保護措施。瞭解研究中使用的措施的可靠性和有效性及其對變化的敏感性至關重要。Godi 等人（2013年）比較了兩種平衡測量——Mini-BESTest 和 Berg 平衡量表——它們在檢測變化方面的敏感性。他們發現 Berg 平衡量表具有更大的天花板效應，這表明 Mini-BESTest 可能是在檢查干預改善平衡效果的研究中使用的更好工具。

測試人員培訓還可以防止檢測威脅。如果使用多個測試者，則應在測試者之間建立評估者間的可靠性。電子和機械措施應接受必要的維護和校準。例如，聽力學家特別認識到校準的重要性，因為校準不佳的儀器會嚴重影響聽力障礙的測試。

實驗者和參與者偏見威脅

當研究過程本身有意或無意地影響研究結果時，就會引入實驗者偏倚。提供干預的人可能會引入實驗者偏倚。

當研究人員為干預組和對照組設定不同的期望時，就會出現經典的實驗者偏差，稱為羅森塔爾效應或皮格馬利翁效應。羅森塔爾效應一詞來自 Rosenthal 和 Jacobson（1968）的一項涉及教師的實驗。羅森塔爾告訴一些教師，他們應該預期學生的智力會有強勁的井噴式增長，而其他教師則沒有得到這些資訊。當老師期望學生表現得更好時，學生的表現會更好。本研究旨在研究這種效果，但當干預領導者傳達對干預組更好結果的期望時，同樣的現象可能會無意中發生。更高的期望成為

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.

自我實現的預言。也許領導者提供了更多的關注或熱情，或者更加努力地提供干預，或者參與者接受了領導者的期望並以同樣的方式回應。

僅僅被分配到一個特定的組可能會引入實驗者偏見。例如，在沒有領導者的提示的情況下，控制參與者可能希望補償沒有被選中進行干預。在許多康復研究中，對照組接受標準治療或「照常治療」。如果對照組意識到干預組正在接受新的東西，他們可能會嘗試補償這種差異。

實驗者可以引入的另一個偏差是治療的補償性均衡。在這種情況下，對照組的干預領導者可能會覺得有必要更加努力地工作，以彌補對照組沒有接受干預的事實。這種類型的偏差類似於 Rosenthal 效應，但針對對照組。對照組也可能向另一個方向做出反應，因為他們沒有接受干預而感到氣餒。作為回應，他們可能不會那麼努力或放棄。這種對有效性的威脅被稱為補償性士氣低落。當對照組和治療組的領導者和/或參與者相互互動時，更有可能發生補償性平等和士氣低落的威脅。

當參與者參與研究影響結果時，參與者偏見威脅就會發揮作用。

當參與者對他們參與研究而不是實際干預這一事實做出反應時，就會發生霍桑效應（Mayo，1949）。該術語來自在 Hawthorne發電廠進行的研究。研究了許多變數以確定哪些因素可能會影響生產力，例如照明或工作站。無論研究什麼，變化多麼微不足道，生產力都會發生變化。得出的結論是，變化而不是實際情況導致了更高的生產力。霍桑效應的發生可能是因為參與者的行為符合預期或想取悅研究人員。

在一項有趣的研究中，銀杏葉對阿爾茨海默病的療效，McCarney 及其同事（2007）將隨訪作為受霍桑效應影響的混雜變數進行了檢查。一些受試者的隨訪很少，而另一些受試者則進行了密集的隨訪。換句話說，強化隨訪組受到了研究人員的更多關注。結果表明，接受強化隨訪的參與者比接受最少隨訪的參與者有更大的認知改善。這一結果是霍桑效應的一個特別顯著的例子，因為認知是使用客觀的標準化評估（ADAS-cog）來衡量的，其中包括 11 項認知任務，例如單詞回憶、定向和遵循命令的能力。

防止實驗者和參與者偏差威脅

與之前許多對有效性的威脅不同，隨機分配到干預組和對照組並不能防止實驗和參與者偏倚。然而，對干預領導者和參與者的盲法提供了針對這些威脅的有力保護。如果領導者和參與者不知道他們是否在提供或接受干預，則更難引入偏倚。這是藥物試驗中的一種常見方法，其中使用安慰劑代替實際藥物。

在康復研究中，通常很難對干預領導者和參與者實施盲法;治療師會知道

現實世界中的證據

缺乏盲法參與者如何導致補償性均等化

在康復和治療實踐中，很難而且在許多情況下不可能使干預領導者和參與者對接受實驗干預的群體不知情。如果你是干預領導者，你必須知道你在領導什麼（而不是提供安慰劑藥丸）;如果您是參與者，您可能會知道您正在參與什麼干預措施。

在現實生活中的例子中，我正在為患有精神障礙的人實施減肥計劃。參與者被隨機分配到干預組或無治療對照組。在知情同意過程之後，參與者被告知他們被分配到哪些組。一些對照組參與者表示希望向研究人員展示他們可以自己減肥，這顯然是對沒有被分配到干預組的補償。在某些情況下，它奏效了，確實有幾名對照參與者在參與研究期間成功減肥。

為了控制這種混淆，兩組都接受某種形式的干預會有所幫助，這樣兩組都不會覺得有必要根據他們的小組作業來證明什麼。作為證據的讀者，通常很難辨別個人何時對實驗者或參與者的偏見做出反應;但是，瞭解本節中討論的保護措施是為了防止潛在問題而實施的，這很有用。

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

他們正在提供干預，並且通常會知道他們何時提供實驗性干預。在一些康復研究中，當對照組接受同等關注的干預時，會提供一種形式的安慰劑。當新的干預措施與標準療法進行比較時，當兩組接受相同的干預時間時，實驗者和參與者的偏倚就不那麼值得關注了。因此，對各組的同等關注通常比無治療對照更可取。但是，參與者可能通過知情同意程式知道他們正在接受實驗性干預。通常，當個人自願參加一項研究時，他們希望接受新的干預，因此如果他們沒有被分配到實驗條件，這可能會導致失望。

可以使用其他方法來最大限度地減少實驗者和參與者的偏倚。明確的干預方案可以減少偏倚。在某些情況下，治療師可能被期望遵循協定，理想情況下不知道哪種干預措施預期會產生更好的結果。限制干預領導者之間的互動以進一步防止偏倚的發展也很有說明。同樣，將干預組和對照組的參與者分開可以減少接受治療的受試者的偏倚。此外，通常可以對進行結果評估的個體實施盲法，以減少或消除評估評分中的偏倚。

流失/死亡威脅

由於損耗引起的威脅，也稱為死亡率，涉及參與者在研究過程中的退出或損失。知情同意的過程承認參與者可以隨時退出研究。個人退出研究的原因有很多，這可能與研究本身有關，也可能無關。個人可能會搬家或遇到其他需要退出的個人問題。其他人可能會退出，因為他們不再對研究感興趣，覺得時間投入太多，或者對干預感到失望。當人們退出一項研究時，它會影響一開始就實現的群體平等化。當大量參與者退出一項研究時，可能會出現群體差異，從而使研究結果變得混亂。

當發生流失時，重要的是要確定退出研究的個體的任何特徵，這些特徵可能使他們與留在研究中的個體不同。也許輟學的人正在經歷更嚴重的情況，在這種情況下，您將不知道干預是否對那群人有效。流失也可能導致組中的參與者人數不偶數。

防止人員流失/死亡威脅

根據研究的時長和對參與者的訪問情況，研究人員可能能夠招募額外的參與者來替換退出的個人，從而保持研究的整體力量。此外，提醒電話和電子郵件等策略可用於促進干預或治療後續的良好出勤率。

應將退出個體的特徵與留下的個體的特徵進行比較。如果存在差異，則應將此因素確定為研究的局限性。統計程式可用於解釋損耗/死亡率威脅，例如對缺失數據使用估計值，但這種方法不如實際參與者評分可取。可以使用「意向治療」分析，其中未接受干預的個體的數據仍包含在分析中。這種分析類似於現實生活中的實踐，在現實生活中，有些人沒有完成或跟進他們治療的所有方面。它還保持了隨機化過程的完整性和組的基線相等性。

練習 5-2

檢測研究中對內部效度的潛在威脅（LO3 和 LO4）

分析以下兩個研究摘要，並確定可能存在哪些對內部有效性的威脅（在提供的選項中）。在查看答案之前，請寫下您認為或不認為特定威脅可能會混淆結果解釋的理由。換句話說，威脅是否意味著干預以外的其他因素導致了改善？您可以在本章末尾找到答案。

研究 #1

Hohler， A. D.， Tsao， J. M.， Katz， D. I.， Dipiero， TJ， Hehl， CL， Leonard， A.， . . .埃利斯，T.（2012 年）。住院運動障礙計劃對非典型帕金森病患者的有效性。帕金森病（2012），871-974。doi：10.1155/2012/871974（Epub 2011 年 11 月 10 日）。

抽象

本文調查了住院運動障礙計劃對非典型帕金森病患者的有效性，這些患者通常對藥物干預反應不佳，並且難以作為門診患者進行康復。91 名非典型帕金森病患者參加了住院運動障礙計劃。患者接受 3 小時 / 天、每周 5 至 7 天的物理、職業和言語治療，並根據日常觀察和數據進行藥物調整。分析了功能獨立性測量（FIM）、計時和開始測試（TUG）、兩分鐘的入院和出院分數之間的差異

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

步行試驗（TMW）、Berg 平衡量表（BBS）和手指敲擊試驗（FT）均顯示出院時有顯著改善（P > .001）。74% 的患者 FIM 總分有明顯的臨床顯著改善。10 名藥物未調整的患者的結果相似。非典型帕金森病患者受益於住院跨學科運動障礙計劃，以改善功能狀態。

考慮一下：

- 本摘要中不包括治療時間。受試者的住院時間從 1 周到 6 周不等，平均為 2.5 周。此外，干預負責人還管理評估。
- 非典型帕金森綜合征癥狀、病程和治療的工作知識將有助於識別有效性的威脅。您可以在 http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001762/ 獲取更多資訊
- 如果您想瞭解有關該研究的更多資訊，您可能需要使用您的圖書館資源來獲取文章的全文。

問題

1. 根據您對摘要和任何其他資源的閱讀，您認為以下哪項對內部有效性的值得注意的威脅？

   A. 成熟
   B. 歷史
   C. 測試

   解釋你的答案。

研究 #2

Frankel， F.， Myatt， R.， Sugar， C.， Whitham， C.， Gorospe， CM， & Laugeson， E.（2010年7月）。一項針對自閉症譜系障礙兒童的父母輔助兒童友誼訓練的隨機對照研究。自閉症與發育障礙雜誌，40（7），827-842。doi：10.1007/ s10803-009-0932-z

抽象

本研究評估了兒童友誼訓練（CFT），這是一種手動的父母輔助干預，旨在提高患有自閉症譜系障礙的二至五年級兒童的社交技能。與延遲治療對照組（DTC）進行比較。目標技能包括對話技巧、同伴進入技巧、發展友誼網路、良好的體育精神、玩耍約會期間的良好主人行為

以及處理挑逗。在後測中，CFT 組在父母的社交技能和遊戲約會行為指標以及兒童的受歡迎程度和孤獨度指標上優於 DTC 組。在 3 個月的隨訪中，父母測量值顯示較基線有顯著改善。事後分析表明，超過87%的接受CFT的兒童在後測時至少在一項指標上表現出可靠的變化，66.7%在隨訪3個月後表現出可靠的變化。

考慮一下：

- 10 名參與者未完成干預，因此未被納入隨訪數據。
- 下表包含在研究中，比較了基線時的兩組。

兒童友誼訓練（CFT）和延遲治療控制（DTC）條件的樣本特徵

變數	群		p
	CFT M（標清） n = 35	DTC M（標清） n = 33	
年齡（月）	103.2 (15.2)	101.5 (15.0)	ns
年級	3.2 (1.0)	3.4 (1.2)	ns
SES	44.6 (10.6)	50.6 (11.8)	ns
男性百分比	85.7	84.8	ns
百分之高加索人	77.1	54.5	ns
WISC-III 型語言智商	106.9 (19.1)	100.5 (15.7)	ns
ASSQ	22.4 (7.3)	22.0 (9.3)	ns
VABS			
通信	84.3 (20.5)	79.8 (15.3)	ns
日常生活	67.0 (18.2)	62.4 (15.7)	ns

繼續

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.

兒童友誼訓練（CFT）和延遲治療控制（DTC）條件的樣本特徵（續）

變數	群		p
	CFT M（標清） n = 35	DTC M（標清） n = 33	
社會化	66.3 (10.8)	66.1 (10.8)	ns
複合	68.1 (16.4)	64.4 (11.0)	ns
# 工作階段參加	11.3 (0.8)	10.7 (1.9)	ns

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

如果您想瞭解有關該研究的更多資訊，您可能需要使用您的圖書館資源來獲取全文文章。

問題

2. 根據您對摘要和任何其他資源的閱讀，您認為以下哪項對內部有效性的值得注意的威脅？

A. 成熟
B. 選擇
C. 檢測
D. 減員

解釋你的答案：

外部 有效性

外部效度是研究結果可以應用於其他人和其他情況的程度。外部效度說明瞭研究的普遍性。當一項研究反映現實世界的實踐時，它具有更多的外部效度。作為循證從業者，那些包括與您的實踐更相似的條件的研究將具有更多的外部有效性，並且可以更有信心地應用結果。

對外部有效性的威脅

當情況或研究參與者與現實世界或臨床環境不同時，就會對外部有效性構成威脅。與內部效度一樣，外部效度是一個連續體;一項研究的外部效度可能好或壞，但它永遠不會完全有效。作為一名循證從業者，評估研究中人員和情況的特徵以確定這些特徵與您自己的實踐有多相似，這一點很重要。表 5-3 總結了對外部有效性的威脅和針對這些威脅的保護措施。

表 5-3 對外部有效性的威脅及其保護

威脅	影響泛化性的混雜因素	保護
採樣誤差	• 樣本不代表總體。	• 使用隨機分配。 • 使用大樣本。 • 從多個網站選擇參與者。• 使用新樣本複製研究。
生態差 有效性	• 研究條件與現實世界的實踐大不相同（當研究以與現實生活實踐密切相關的方式進行時，結果將更具普遍性）。	• 確保研究人員對實際實踐問題敏感。 • 通過有效性研究進行複製。

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

採樣誤差

定量研究的一個主要原則是將樣本的結果推廣到總體中。抽樣誤差是總體和樣本之間存在的任何差異。抽樣誤差的確切性質並不總是已知的，因為總體的許多特徵是未知的。然而，在已知特徵中，可以將樣本與總體進行比較，以確定相似之處和不同之處。例如，男孩被診斷出患有自閉症的可能性大約是女孩的五倍（CDC，2012），儘管即使這是來自樣本的估計。在一項旨在代表自閉症兒童的研究中，更具代表性的樣本將是具有相似性別分佈的樣本。

防止採樣誤差

抽樣方法會影響外部效度。儘管獲得代表性樣本的最佳方法是從目標總體中隨機選擇大樣本，但這並不總是可行的。在隨機抽樣中，總體中的每個個體都有相同的機會被選中。使用大型隨機樣本，您可能會選擇一組代表總體的參與者。不幸的是，真正的隨機抽樣在醫療保健研究中很少發生，因為從整個人群中抽樣通常是不切實際的。例如，在對多發性硬化症患者的干預研究中，人群將是所有多發性硬化症患者。在全球範圍內進行採樣然後進行干預幾乎是不可能的。

真正的隨機抽樣確實發生在一些研究中，當樣本較小且可獲取時。例如，可以通過對成員名單進行隨機抽樣來獲得對美國職業治療協會成員的研究。

醫療保健研究中最常見的抽樣方法是便利抽樣。在這種方法中，選擇參與者是因為研究人員很容易獲得他們。在對患有特定病症或殘疾的個體進行研究時，研究人員很可能會去為這些個體提供服務的治療機構。然後，研究人員可能會要求志願者，或者臨床醫生可能會在符合研究標準的人入院時接近該人。過程中缺乏隨機性很有可能引入偏差或抽樣誤差。例如，當從一所學校、一個社區或一個診所中選擇樣本時，他們更有可能具有與整個人口不同的特徵;根據環境，他們可能更貧窮、年齡較大或癥狀更嚴重。

減少採樣誤差的一種方法是從多個設置中選擇較大的樣本。樣本越大，越有可能近似於總體。此外，還可以選擇多個設置來表示總體的異質性。例如，在考慮對患有注意力缺陷多動障礙（ADHD）的兒童的研究的普遍性時，如果從該國不同地區的城市和郊區招募兒童，樣本更有可能代表人口，以更好地代表人口的種族和社會經濟特徵。

在研究的結果部分，研究人員必須提供研究參與者的詳細描述。許多期刊要求至少包括性別、年齡和種族。作為循證從業者，您可以查看此資訊以確定樣本是否代表總體。但是，對您來說更重要的是研究中的樣本是否與您合作的客戶相似。當研究樣本與您的客戶相似時，您更有理由概括研究結果。

生態有效性威脅

生態效度是指研究發生的環境以及它與現實世界的接近程度。進行研究的治療方法或方法、研究發生的時間以及研究的發生地點都是影響研究外部有效性或普遍性的重要考慮因素。有時，研究中干預的管理人員比典型的從業者訓練有素。研究時間可能比大多數保險公司承保的住院時間長，或者干預可能比標準做法更強烈。該研究可能在住院環境中進行，儘管大多數患有這種疾病的人實際上是在門診接受治療。與研究條件和實際實踐的任何差異都代表著對外部有效性的威脅。特定研究的泛化性在與研究中相似的情況下會很好，而在不同的情況下會很差。

防止生態有效性威脅

從業者更有可能應用與現實世界實踐相關且實用的研究，並且當研究對典型的實踐情況敏感時，研究具有更大的生態效度。例如，研究人員可以確保干預在客戶接受治療的典型時間範圍內進行，或者提供干預的治療師是已經在特定類型醫院工作的人。

作為從業者，重要的是要謹慎應用研究結果並考慮與您的相似性

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.

自己的情況。當研究的特徵相似時，研究更易推廣到您的實踐和客戶。例如，Sutherland 等人（2012）的一項研究考察了退伍軍人創傷后應激障礙（PTSD）的暴露療法，將更適用於涉及治療患有 PTSD 的退伍軍人的實踐情況。該研究的結果在治療遭受過性虐待的女性的 PTSD 方面的適用性較差，外部效度也較差。在另一個例子中，一項涉及 24 周太極拳的設計良好的研究表明，它能有效改善帕金森病患者的平衡（Tsang，2013 年）。但是，如果您在 24 周的時間內無法見到客戶，則這項研究與您的實踐環境不太相關。儘管如此，您可以使用這項研究的結果向您的管理人員和/或保險公司證明為什麼需要更長的逗留時間。

通過複製提高泛化能力

複製或可重複性對於研究的推廣至關重要，也是科學方法的主要原則。它在樣本和情境的泛化中都很重要。研究結果必須能夠重複，以確保結果的普遍性和適用性。如果幾項研究對干預的有效性或結果的預測指標得出相似的發現，臨床醫生可以對這些結果更有信心。

複製的另一個考慮因素是研究人員本身。即使有幾項研究支持某種特定方法，如果所有這些研究都是由同一位研究人員進行的，也應該有一些擔憂

結果不會推廣到其他情況。一位研究人員能夠比另一位研究人員獲得更多積極的發現可能是有原因的。也許那個研究人員和他或她的團隊是傑出的治療師，是他們的一般能力而不是實際的干預造成了差異。

此外，當發現特別令人驚訝或顯著時，複製很重要。這類發現很有趣，因此被發表的可能性更高。但是，複製將揭示發現是否是偶然的結果（即 I 類錯誤）。

複製通常是一個程度問題。研究很少遵循先前研究的確切程序來確定是否獲得相同的結果。通常，變數縱以擴展先前研究的結果。重複研究可以縮短乾預期，使用不同的結果測量，將方法應用於不同的樣本，或在新的環境中進行干預。研究建立在先前工作基礎上的能力是科學方法力量的一部分。

內部 對 外部 有效性

在設計研究時，研究人員必須在內部和外部有效性之間找到平衡。嚴格控制以最大化內部效度的研究將具有較低的外部效度。例如，產生同質樣本的納入標準、管理干預的嚴格方案、專家干預

現實世界中的證據

複製如何改變對促進通信的看法

在 1990 年代初期，人們對促進自閉症患者的溝通產生了濃厚的興趣。這種興趣大部分來自 Biklen 及其同事（1992 年）的工作，他們使用這種技術獲得了令人驚訝的結果。在促進溝通中，促進者提供身體幫助，說明自閉症患者在鍵盤上打出資訊。假設是促進溝通克服了干擾自閉症患者溝通能力的神經運動困難。

在一項針對 43 名自閉症患者的非對照研究中，Biklen 及其同事報告了令人震驚的結果。以前，非語言個體會寫語法正確的句子和段落，甚至詩歌。對這些發現的懷疑導致其他研究人員進行了對照研究。Green（1994）的一篇評論發現，當促進者的影響力受到控制時，該技術就不再有用。審查表明，引導者相信促進溝通的潛力和客戶尚未開發的能力，導致他或她無意識或無意地指導溝通過程。現在，美國言語和聽力協會和美國心理學會等組織斷言，沒有證據支援自閉症患者促進溝通。

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?

領導者和有限地接觸替代療法將產生可以在因果關係背景下解釋的結果，但不能反映日常實踐。相比之下，在現實世界條件下進行的研究「更混亂」;也就是說，沒有那麼多的控制措施，現實世界的條件引入了對結果的更多替代解釋——以犧牲內部效度為代價來提高外部效度。

在積累有關特定干預措施的研究證據時，您可以使用一個過程，從內部效度高的研究開始，然後轉向外部效度更高的研究。首先，重要的是要知道干預措施在理想條件下是否有效;也就是說，一項具有很強內部效度的高度對照研究。

一旦確定了干預的有效性，未來的研究就可以在更典型的實踐條件下檢查相同的干預。這些研究之間的差異可以稱為療效與有效性。療效研究是強調內部有效性並檢查干預措施在理想條件下是否有效的研究。通過療效研究，您可以更有信心地認為干預措施是產生差異的原因;但是，研究的條件可能與現實世界的條件不同。

在有效性研究中，研究條件更能反映現實世界的實踐;然而，練習的不整潔性質意味著遊戲的內部有效性可能會面臨更多威脅。關於治療實踐的研究總是會對有效性構成威脅。研究人員在設計研究時面臨重大挑戰，必須找到一種平衡，包括最大限度地減少威脅、務實和合乎道德地運營。從證據 5-2 提供了一個有效性研究的例子，該研究在現有的社區健身設施中進行力量訓練干預。

一位研究人員對研究一種新的干預措施感興趣，該干預措施旨在增加小學 1 至 3 年級兒童的體育活動量和強度。兩所學校已同意參與這項研究。一所學校位於城市環境中，學生大多來自較低的社會經濟和種族多元化背景。另一所學校位於郊區，學生大多來自高社會經濟背景和高加索人背景。

問題

考慮以下問題，並描述研究人員如何減少對有效性的威脅。以下情況同時解決了內部和外部有效性問題。

1. 研究人員計劃實施該研究的學校不允許研究人員將兒童隨機分配到小組中。有效性受到的威脅是什麼，研究人員如何管理這種威脅？

2. 研究人員計劃將攀岩牆和其他新的但昂貴的設備作為活動計劃的一部分，從而提高人們對體育活動的興趣。有效性受到的威脅是什麼，研究人員如何管理這種威脅？

3. 為了確定學校的活動是否延續到家裡，要求家長記錄孩子參與活動的一周，其中包括活動類型、參與的時間和強度水準。有效性受到的威脅是什麼，研究人員如何管理這種威脅？

練習 5-3

管理特定研究中對有效性的威脅
（LO3、LO4、LO5 和 LO6）

兒童肥胖是一個重大的公共衛生風險，已經做出了許多努力來解決這個問題

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?*

來自證據 5-2

有效性研究示例

Minges， KE， Cormick， G.， Unglik， E.， & Dunstan， DW （2011）.成人糖尿病患者或有患糖尿病風險的阻力訓練計劃的評估：社區環境中的有效性研究。國際行為營養與身體活動雜誌，8,50。doi：10.1186/1479-5868-8-50.

注 A：研究人員有興趣在受控條件下採取有效的干預措施，並評估其在現有健身設施中的有效性。

注 B：在健身中心，大量退出（86 個月時有 2 名參與者，但 6 個月時只有 32 名參與者）並不意外。人們經常停止他們的健身計劃。

背景：
研究基於社區的阻力訓練計劃 （Lift for Life） 對患有 2 型糖尿病或有患 2 型糖尿病風險的成年人的腰圍和功能測量的影響。

方法：
Lift for Life 是一項從研究到實踐的計劃，旨在為患有 2 型糖尿病或有患 2 型糖尿病風險的成年人提供循證阻力訓練計劃，以傳播到澳大利亞社區現有的健康和健身設施。對在澳大利亞墨爾本的 4 個活躍提供者中訪問該計劃的 86 名參與者進行了回顧性評估。這項縱向研究的主要目標是評估基於社區的阻力訓練計劃的有效性，從而排除隨機、對照的研究設計。在基線時收集腰圍、下半身 （椅子坐姿到站姿） 和上半身 （手臂捲曲試驗） 力量和敏捷性 （定時起跳） 測量，並在 2 個月 （n = 86） 和 6 個月時重複 （n = 32）。

結果：
相對於基線，平均腰圍 （-1.9 cm， 95% CI： -2.8 至 -1.0） 和定時敏捷性測試 （-0.8 秒，95% CI： -1.0 至 -0.6） 顯著減少;8 周結束時，下半身 （重複次數： 2.2， 95% CI： 1.4-3.0） 和上半身 （重複次數： 3.8,95% CI： 3.0-4.6） 力量顯著增加。在16周的評估中仍然存在顯著差異。在基線和 24 周隨訪時進行完整測量的 32 名參與者的年齡和性別調整的匯總時間序列回歸分析顯示，腰圍和功能測量有顯著的時間影響，在 24 周評估中觀察到與基線相比的變化最大。

結論：
這些發現表明，在社區環境中為患有 2 型糖尿病或有患 2 型糖尿病風險的人進行的循證阻力訓練計劃可以帶來良好的健康益處，包括減少向心性肥胖和改善身體機能。

注 C：如果沒有對照組，您可能不太確定 Lift for Life 是否產生了影響。也許參加健身中心的人利用了其他計劃，或者更有可能在計劃之外鍛煉。因為評估員不會對小組分配視而不見，所以他們可能會有意識或無意識地在對他們希望改進的個體進行評分時表現出偏見。此示例演示了提高外部有效性有時如何損害內部有效性。

FTE 5“2 問題　　以 Lift for Life 研究為例，提高外部效度如何損害內部效度？

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.

危急 思維 問題

1. 為什麼在研究中，大樣本通常比小樣本更可取（至少給出三個原因）？

2. 為什麼隨機對照試驗被認為是最強的單一研究設計？

3. 為什麼隨機分配到小組會導致道德問題？

4. 儘管預測試通常是可取的，但它們如何可能對有效性構成威脅？

5. 對於以下三種情況中的每一種，研究人員如何管理對有效性的威脅以確定新的干預措施是否有效？

   - 新干預與無治療對照組的比較
   - 將新干預措施與常規治療對照組進行比較
   - 將一種新的干預措施與另一種循證干預措施進行比較

6. 解釋隨機選擇和隨機分配之間的區別。這些研究實踐涉及有效性的哪些方面，以及如何解決？

7. 為什麼很難設計一項內部和外部效度都很強的研究？您如何平衡這兩種類型的效度？

答案

運動 5$1

1. 存在漁業威脅的原因有兩個：研究人員沒有研究假設，並且正在研究四個結果。

2. 如果研究人員有關於哪些矯形器最適合哪些結果的先前假設，那麼這項研究會更有力。這可能基於現有研究或研究人員的臨床經驗。為了解決研究多個結果的事實，研究人員應調整統計分析的alpha水準或使用統計數據來控制多重比較。

3. 將10人分成三組將導致一項功效非常低的研究。

4. 研究人員將希望招募更多的參與者，並且可能需要使用另一家診所或在更長的時間內進行研究。也可以通過減少組數來增加功效，因此研究人員可以比較兩種矯形器（儘管每組最好還是有5名以上的參與者）或使用交叉設計，其中所有參與者都嘗試所有矯形器。

運動 5$2

1. 研究 #1 A. 成熟—否。雖然沒有對照組並且研究持續了數周，但在確定成熟是否對有效性構成威脅時，請考慮疾病的正常過程。帕金森病的正常病程正在逐漸惡化，因此如果不治療，您不會期望病情會有所改善。

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

B. 歷史——是的。該研究的結論是，跨學科運動障礙計劃可有效改善帕金森病患者的運動問題。藥物調整可能是一種歷史威脅：91 名參與者中有91名在干預期間接受了藥物調整。儘管沒有調整藥物的參與者有類似的改善，這為治療提供了一些支援，但這項研究的設計使得很難確定是治療還是藥物（或兩者）造成了差異。

C. 測試 — 是的。測試存在一些問題。

Timed Up and Go 測試使用時間作為結果，而 2 分鐘步行測試則根據行駛的距離來衡量。有了這些客觀結果，您就不會那麼擔心有偏見的評估。然而，FIM 和 Berg 平衡量表確實涉及治療師的判斷，參與干預並希望看到改善的治療師可能傾向於將參與者的評分（儘管是無意的）高於無偏倚的評分者。此外，帕金森病的藥物有效性在一天中有所不同，因此進行評估的時間可能會影響結果。

2. 研究 #2 A. 成熟——不，在這種情況下，有一個隨機分配的無治療對照組。如果對照組有改善，則干預的改善大於對照組的改善，這表明干預組的改善超過了任何典型發展。B. 選擇 - 否，隨機分配有助於促進平等的組分配。該表提供了額外的支援，即兩組在開始時具有可比性。C. 檢測 - 是。自我報告、家長報告和教師報告有問題。孩子、家長和老師都知道干預，並且可能偏向於提供更積極的報告。使用觀察方法（即在社交場合觀察孩子）將增強研究的這一方面。D. 流失——是的。干預組的10名兒童在隨訪測試期間不可用： 其中8名兒童退出，2名兒童因行為原因被移走。這10名兒童可能對干預的反應不佳，這可能就是他們退出的原因。被撤職的兩人並沒有得到好處。如果這些兒童被納入研究結果，結果可能會（甚至很可能）不那麼積極。這應該被考慮在

在評估干預的有效性和數量時。

運動 5$3

1. 缺乏隨機化的主要問題是選擇的有效性會受到威脅。為了解決這個問題，使用可以減少組之間可能發生的任何差異的策略非常重要。在基於學校的研究中，一個教室接受干預而另一個教室沒有干預是很常見的。您不希望將一所學校作為干預設置，將一所學校作為控制設置，因為學校的明顯差異可能是您在干預中發現的差異的原因。相反，您可以隨機分配每所學校的教室來接受或不接受干預。您可以通過使用候補名單控制設計來解決對照組未接受干預的道德問題（即，您最終將向對照組提供干預）。這種方法的一個缺點是可能存在更大的實驗者和參與者偏倚。對測試人員實施盲法並減少學生和教師的接觸，尤其是在體育活動期間，將有助於解決這些問題。在不引入額外體育活動的情況下，為對照組提供對新事物的同等關注也很有用。例如，您可以讓對照組參與棋盤遊戲。

2. 包括昂貴的設備使得其他學校不太可能實施這種干預，從而使其不太普遍。研究人員應考慮重新設計干預措施，以使用大多數學校環境中通常可用的設備;但是，這樣做可能會使研究人員失去新設備所帶來的新奇感或興奮感。

3. 要求父母保留日誌會引入檢測威脅。維護一周的日誌對父母的要求很高，而且您不太可能收到完整的數據。研究人員可以使用更客觀的方法，例如孩子們佩戴的加速度計來記錄參與活動所花費的時間。另一種對父母來說負擔較小的方法是時間採樣。在時間採樣中，通常以隨機間隔，計時器指示應進行日誌輸入。使用此方法，父級只需回應計時器，而不必一直保留記錄。這兩種方法仍然存在問題。例如，父母可能忘記打開加速度計，孩子可能會丟失加速度計，或者父母可能仍然對時間採樣方法沒有反應。

所有這些例子都說明了設計研究的挑戰。設計一個完美的

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.

沒有有效性威脅的研究。研究人員通常會權衡他們的選擇，並根據特定的研究問題、提出的道德問題和實用問題做出選擇。

證據 5$1

1. 是的，所有比較的 p 值均> 0.05，表明基線時各組之間沒有統計學上的顯著差異。

2. 如果各組開始時背痛程度不同，這可能會影響/混淆研究結果。例如，如果對照組的疼痛較少，而干預組在基線時疼痛較多，即使沒有干預，成熟也可能導致干預組有更大的恢復，因為干預組可能有更多的改進空間。對照組可能無法改善太多，因為他們已經沒有經歷太多的痛苦。在這個證據的例子中，兩組在背痛的結果測量以及其他人口統計變數上是等效的，這是一件好事。

證據 5$2

如果沒有對照組，您可能不太確定 Lift for Life 是否產生了影響。也許參加健身中心的人利用了其他計劃，或者更有可能在計劃之外進行鍛煉。此外，您預計評估員的精確度會降低，他們不會是盲人，並且可能因網站而異。

引用

Biklen， D.， Morton， M. W.， Gold， D.， Berrigan， C.， & Swaminathins， S.和Teng，EJ （2012）。針對患有 PTSD 的退伍軍人為期 12 周的基於群體的暴露療法模型的初步研究。創傷與壓力雜誌，25（2），150-156。
（1992）. 促進溝通：對自閉症患者的影響。語言障礙主題，12（4），1-28。

曾偉文 （2013）。太極拳訓練可有效減少帕金森病患者的平衡障礙和跌倒。物理治療雜誌， 59， 55.

疾病控制和預防中心 （CDC）。（2012）. 自閉症譜系障礙的患病率：自閉症和發育障礙監測網路，14 個網站，美國，2008 年。取自 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， JC，Hill， J.， & Gusi， N.（2012）.一種基於互聯網的職業干預，用於預防亞急性腰痛的慢性病：一項隨機對照試驗。康復醫學雜誌，44,581-587。Frankel，F.， Myatt， R.， Sugar， C.， Whitham， C.， Gorospe， CM， & Laugeson， E.（2010年7月）。一項針對自閉症兒童的父母輔助兒童友誼訓練的隨機對照研究

譜系障礙。自閉症與發育障礙雜誌，
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）.Mini-BESTest 和 Berg 平衡量表在平衡障礙患者中的信度、效度和反應性比較。物理治療，93,158-167。格林，G.（1994 年）。促進者的影響力：促進者的品質

dence 的在 HC Shane（編輯）中，促進溝通：臨床和社會現象（第 157-226 頁）。加利福尼亞州聖地牙哥：單數。

基倫，JD，福特曼，SP，紐曼，B.和瓦拉迪，A.（1990 年）。評估將尼古丁口香糖與自我指導行為治療相結合的預防吸煙復發的治療方法。諮詢與臨床心理學雜誌，58,85-92。

梅奧，E.（1949 年）。霍桑和西部電氣公司：工業文明的社會問題。英國倫敦：勞特利奇。

麥卡尼，R.，華納，J.，伊利夫，S.，范哈塞倫，R.，格裡芬，M.和費舍爾，P.（2007年7月3日）。霍桑效應：一項隨機對照試驗。BMC 醫學研究與方法，7,30。Minges， KE， Cormick， G.， Unglik， E.， & Dunstan， DW （2011年5月25日）。成人糖尿病患者或有患糖尿病風險的阻力訓練計劃的評估：社區的有效性研究

nity 設置。國際行為營養與物理雜誌
活動，8,50。doi：10.1186/1479-5868-8-50 羅森塔爾，R.，和雅各森，L.（1968）。教室里的皮格馬利翁。紐約州紐約：霍爾特，萊因哈特和溫斯頓。Sabbag， S.， Twamley， E. W.， Vella， L.， Heaton， R. K.， Patterson，T. L.， & Harvey， P. D.（2012）.精神分裂症患者日常功能自我評估準確性的預測因數。精神分裂症研究，137,190-195。

薩蒙斯·普雷斯頓。（日期不詳）。Jamar 手動測功機用戶手冊。取自 https://content.pattersonmedical.com/PDF/spr/ Product/288115.pdf