CHAPTER 9

Designing Experiments

9.1. Variables in Studies

Chapter 8 was about *who* we study: how to select subjects using sampling methods. This chapter is about *how* we study them and what conclusions we can draw from different designs.

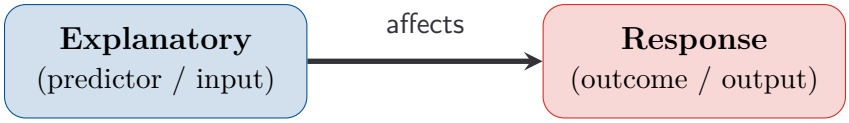
We will follow one study through the first half of the chapter: a comparison of **Ozempic** (a GLP-1 weight-loss medication) to an **intensive lifestyle programme** (diet + exercise). Each section revisits this study to show how design decisions affect the conclusions we can draw.

? If a study finds that people who take Ozempic lose more weight than people who diet and exercise, does that mean Ozempic *caused* the extra weight loss? What would it take to make that claim?

We first saw explanatory and response variables back in Chapter 5. The concept of these variables extends beyond the setting of linear models and is fundamental to experimental design.

Definition 9.1 *Explanatory and Response Variables*

- The **explanatory variable** is the variable used to explain, predict, or account for changes in the outcome.
- The **response variable** is the outcome we measure or observe.



Example 9.2 *Ozempic vs. Lifestyle Changes*
Context: Researchers study whether Ozempic or an intensive lifestyle programme leads to greater weight loss over 6 months.

- **Explanatory variable:** Type of intervention (Ozempic or lifestyle programme).
- **Response variable:** Weight loss (kg or % body weight) over 6 months.

Example 9.3 *Identify the Variables*
For each scenario, identify the **explanatory** and **response** variable:

1. Researchers randomly assign 80 students to sleep either 6 or 8 hours the night before an exam, then compare their scores.

Explanatory: _____ Response: _____

2. An agricultural station randomly assigns 60 plots to receive a new fertilizer or the standard one, then measures yield at harvest.

   Explanatory: _____ Response: _____

9.2. Observational Studies vs. Experiments

Consider two versions of the Ozempic study from Example 9.2:

Version A: Researchers survey 500 adults on a weightloss journey. Some already take Ozempic; others follow a lifestyle programme on their own. Researchers compare the average weight loss between the two groups.

Version B: Researchers recruit 500 adults and *randomly assign* half to receive Ozempic and half to follow a lifestyle programme. After 6 months, they compare weight loss.

Both versions compare the same interventions and measure the same outcome, but they support very different conclusions because the study designs differ.

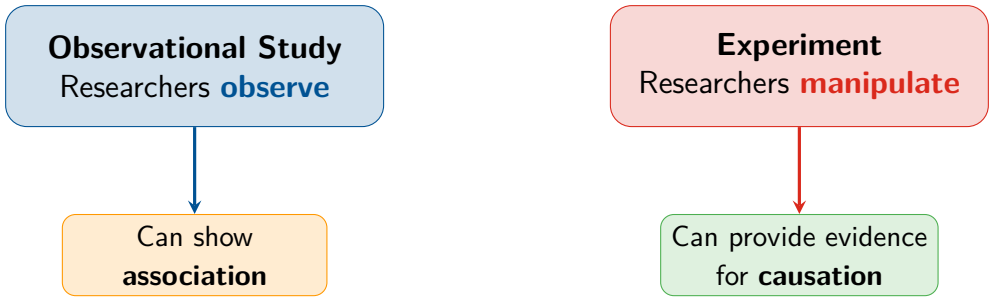
> **Definition 9.4** *Observational Study*
> In an **observational study**, researchers observe and record data without manipulating any variable.

Version A is observational: no one was told which treatment to use. Researchers simply recorded what people were already doing.

> **Definition 9.5** *Experiment*
> In an **experiment**, researchers deliberately impose a treatment on subjects and measure the response.

Version B is an experiment: researchers decided who would receive Ozempic and who would follow the lifestyle programme.

The diagram below shows the key difference. The question to ask is always: did the researcher *manipulate* the explanatory variable, or merely *observe* it?



> **Key point:** Observational studies can show that two variables are **associated**, but cannot establish causation.

Example 9.6 *Classify Each Study*
Is each study **observational** or an **experiment**?

1. A researcher records the amount of screen time 200 teenagers report they spend each week and measures their sleep duration.

   _____

2. Students are randomly divided into two groups: one uses flashcards, the other uses practice exams. Test scores are recorded.

   _____

Why Use Observational Studies?

If experiments support causal claims, why not always run them? Sometimes researchers *cannot* randomly assign subjects to groups.

Ethical considerations. Randomly assigning people to smoke or not smoke to study lung cancer would deliberately expose subjects to a known health risk. Whenever the treatment could **cause harm**, research ethics boards will not approve the experiment. Researchers must observe existing behaviour instead.

Feasibility. Randomly assigning people to attend or skip university to study lifetime earnings would require controlling decades of life choices. Some explanatory variables simply **cannot be manipulated** by a researcher: they are fixed characteristics (e.g., age, gender) or personal decisions beyond the study's control.

> **Key point:** When random assignment is impossible, researchers usually have to rely on observational data. On its own, though, an observational study cannot reliably establish causation.

Exercise 9.7 *When Is an Experiment Impossible?*
For each research question, explain whether a randomized experiment is possible. If not, state why and identify what type of study should be used instead.

(a) A researcher wants to know whether childhood exposure to air pollution causes lower IQ scores in adults.

(b) A researcher wants to know whether daily vitamin D supplements reduce the risk of seasonal

depression.

9.3. Confounding Variables

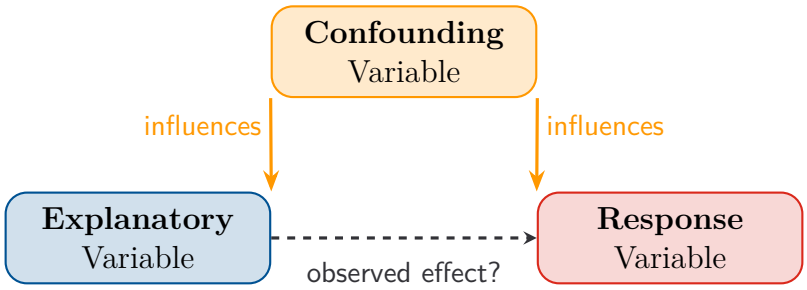
Return to the observational Ozempic study (Version A from Section 9.2). Suppose it finds that Ozempic users lost more weight than lifestyle-programme users. Can we conclude that Ozempic is more effective?

> **Definition 9.8** *Confounding Variable*
> A **confounding variable** (also called a **lurking variable**) is associated with *both* the explanatory and the response variable, making it hard to tell what actually caused the observed effect.

Not necessarily. Confounding variables make the answer far less clear:

- **Income:** Ozempic costs several hundred dollars per month, so users likely have higher incomes. This is linked to better nutrition, gym access, and lower stress.
- **Baseline BMI:** People with more severe obesity may be more likely to seek medication.
- **Motivation and healthcare access:** These also differ between the groups.

Any of these variables could help with weight loss *independently* of the drug itself, making it impossible to attribute the difference to Ozempic alone.



So even if Ozempic had no effect at all, the observational study could still show greater weight loss in the Ozempic group. The groups may differ in income, baseline BMI, or access to care. Any of those differences could explain the result.

Example 9.9 *Confounders in the Ozempic Study*
Context: The observational Ozempic study (Version A) finds that Ozempic users lost more weight than lifestyle-programme users.

Task: Besides income, list at least two other plausible confounding variables.

Example 9.10 *Spot the Confounder*
For each claim, identify a plausible **confounding variable**:

1. "Children with more books at home score higher on reading tests."

   ---

2. "Students who sit in the front row get higher grades."

> ✪ **Common Mistake:** Seeing a strong association and concluding causation. **Association does not imply causation.** Always consider confounders before drawing causal conclusions from observational data.

9.4. The Four Principles of Experimental Design

An experiment only helps if it is designed well. Without good controls, it can be just as misleading as an observational study. Four principles make experiments more reliable.

9.4.1 Principle 1: Comparison

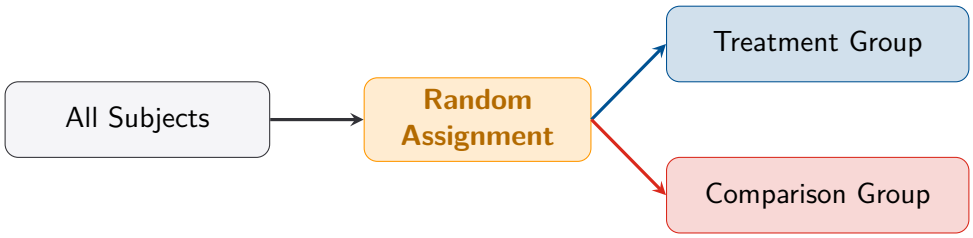
> **Definition 9.11** *Comparison, control group*
> A well-designed experiment needs a **meaningful comparison**. This is typically achieved by having a **treatment group** receiving the intervention and a **comparison group** receiving a different condition.
>
> When the comparison group receives a placebo (a fake treatment designed to look like the real one; defined formally in §9.7), no treatment, or the current standard treatment, we call it the **control group**.

In the Ozempic experiment, the treatment group receives Ozempic and the comparison group follows the lifestyle programme. Without a comparison group, we could not tell whether the weight loss came from the drug or from something else entirely: seasonal diet changes, a new gym opening nearby, or simply the passage of time.

A comparison group is necessary but not sufficient. If the groups already differ before the study begins, the outcome difference might just reflect those pre-existing gaps. That is, where randomization comes in.

9.4.2 Principle 2: Randomization

Definition 9.12 *Randomization*
Randomly assign subjects to treatment and comparison groups. This tends to make the groups comparable *before* any treatment is applied, especially when the sample is large enough (see Replication).



Why must the assignment be *random*? Any systematic rule risks correlating with a confounder. Assigning the first 100 volunteers to Ozempic and the next 100 to the lifestyle programme might seem fair, but early volunteers may be more motivated or more informed about the study. Assigning by clinic location could cluster participants by neighbourhood, income, or ethnicity. Because random assignment introduces no systematic pattern, it tends to distribute confounding variables, including ones we have not thought of, more evenly across groups.

Random assignment also supports statistical inference: it lets us use probability to judge whether an observed difference is likely due to chance (as we saw in Chapter 8).

Exercise 9.13 *Why Not Let Subjects Choose?*
In the observational Ozempic study, people chose their own treatment. Why does this create a problem that randomization solves?

Randomization distributes both known and unknown confounders across groups. But when we already know that a variable strongly affects the response, we can do better than leaving the balance to chance. That is the motivation for blocking.

9.4.3 Principle 3: Blocking

Definition 9.14 *Blocking*
Blocking groups similar subjects together *before* randomly assigning them to treatments. Each group is called a **block**.

Suppose the Ozempic study's 200 participants range from BMI 25 (mildly overweight) to BMI 45 (severely obese). Someone with BMI 45 might lose 15 kg on either programme, while someone with BMI 25 might lose only 3 kg. If randomization happens to put more severely obese participants in the Ozempic group, the drug will look artificially effective.

Blocking solves this: divide participants into BMI categories (e.g., 25–30, 30–35, 35–40, 40+), then randomly assign treatments *within* each block. Now both the Ozempic and lifestyle groups contain similar proportions of each BMI range, and any difference in weight loss is less likely to be explained by BMI imbalance.

> 👁 **Intuition:** Blocking is like stratified sampling, but for experiments. In stratified sampling, we divide the population before *selecting* subjects. In blocking, we divide subjects before *assigning* treatments.

A commonly used mantra for the role of randomization and blocking is:

"Block what you can, randomize what you cannot."

Even with well-balanced groups, a study can still mislead if there are too few subjects. Random variation can easily produce a large observed difference by chance alone. That is what replication prevents.

9.4.4 Principle 4: Replication

Definition 9.15 *Replication*
Use **enough subjects** in each treatment group so that results are not easily explained by chance alone.[1]

If you flip a fair coin 4 times, getting 3 heads is not surprising: that is just what happens with small numbers. But in 400 flips, getting 300 heads would be suspicious.

Small studies are noisy. With only a handful of subjects, random variation can easily mask a real treatment effect or create a false one. Larger groups make it easier to detect real differences and harder to mistake chance variation for a treatment effect.

Replication also matters for generalizability: if an experiment is repeated with different subjects and yields similar results, the findings are less likely to be a fluke. We will not go deeper into that distinction here.

Applying the Four Principles

In practice, a good experiment uses all four principles together. Returning to our Ozempic study:

- **Comparison:** Divide participants into a treatment group (Ozempic) and a comparison group (lifestyle programme): without a comparison group, any change in weight could be due to the passage of time, not the treatment.
- **Randomization:** Randomly assign participants to groups within each block, distributing both known and unknown confounders evenly across groups.
- **Blocking:** Since baseline BMI is known to affect weight loss, group participants into BMI categories (e.g., 25–30, 30–35, 35–40, 40+) before assigning treatments, so each range is

[1] Within the context of our course, replication means using enough experimental units *within* a single experiment, not repeating the whole experiment from scratch (this is the secondary meaning of replication).

represented in both groups.

- **Replication:** Recruit enough participants per group so that results are not easily explained by chance variation.[2]

Each principle handles a different problem: comparison gives you a baseline, randomization balances confounders you have not thought of, blocking removes known sources of variation, and replication keeps random noise from swamping the signal.

Example 9.16 *Remote vs. In-Person Productivity*
Context: A tech company wants to test if remote work (3 days from home) improves productivity, measured by tasks completed per week (standardized by role). They allow employees to choose, and 3 volunteers switch to remote work. The remote group completes 15% more tasks per week than the company average.

Task: Identify at least two problems with this study design, using the four principles as a guide.

Example 9.17 *Redesigning the Study*
Task: The remote-work study (Example 9.16) had no comparison group, no randomization, and only 30 volunteers. Apply the four principles to redesign it.

1. **Comparison:** _____
2. **Randomization:** _____
3. **Replication:** _____
4. **Blocking:** _____

9.5. Factors, Levels, and Treatments

Once the design principles are in place, we also need to describe the *structure* of the treatments. Our Ozempic example only had a single explanatory variable with two possible values (Ozempic vs. lifestyle). Many experiments, however, investigate *multiple* explanatory variables simultaneously, such as medication type *and* exercise intensity. We need vocabulary for that.

[2]The exact number depends on the expected effect size and variability, but generally, more is better. The required counts are beyond the scope of our course and are usually determined by **power analysis**.

Definition 9.18 *Factor and Level*
A **factor** is an explanatory variable that is deliberately manipulated in an experiment. The specific values a factor can take are called its **levels**. A factor can have two levels (e.g., drug vs. placebo) or many (e.g., low / medium / high dose).

Definition 9.19 *Treatment*
A **treatment** is a specific combination of factor levels applied to a subject. If there are k factors with $n_1, n_2, \ldots, n_k$ levels:

$$\text{Number of treatments} = n_1 \times n_2 \times \cdots \times n_k$$

Example 9.20 *Medication and Exercise Intensity*
Context: Researchers want to know whether weight loss depends on the type of medication, the intensity of exercise, or both. They design a study with two factors:

- **Factor 1: Medication** with 3 levels: Ozempic, Metformin, Placebo.
- **Factor 2: Exercise intensity** with 2 levels: Moderate (30 min walk, 3×/week) and Vigorous (45 min run, 5×/week).

Since there are 2 factors with 3 and 2 levels, the total number of treatments is $3 \times 2 = 6$. Each treatment is a unique combination of medication and exercise:

#	Medication	Exercise
1	Ozempic	Moderate
2	Ozempic	Vigorous
3	Metformin	Moderate
4	Metformin	Vigorous
5	Placebo	Moderate
6	Placebo	Vigorous

Each participant is randomly assigned to exactly one of these six treatments. Notice that "Ozempic" alone is not a treatment here; the treatment is "Ozempic + Moderate exercise" or "Ozempic + Vigorous exercise." The treatment is always the full combination of all factor levels.

Example 9.21 *Water Temperature and Fertilizer*
Context: A study tests the effect of water temperature (20°C vs. 30°C) and fertilizer type (A vs. B) on plant growth.

- Factors: _____
- Number of factors: _____
- Levels per factor: _____
- Number of treatments: _____

All treatments:

#	Temperature	Fertilizer
1	20°C	A
2	20°C	B
3	30°C	A
4	30°C	B

Example 9.22 *Counting Treatments*
Task: Determine the number of treatments for each experiment.

1. A study on pain relief uses 3 doses of medication (low, medium, high) and 2 delivery methods (pill, injection).

   Factors: _____ Treatments: _____

2. A marketing study on customer engagement has three factors: ad design (4 versions), colour scheme (3 options), and font (2 options).

   Factors: _____ Treatments: _____

9.6. Blocking and Design Types

Randomisation controls for lurking variables *on average*, but only if the groups it creates are similar. When subjects differ in a way that is known to affect the response, pure randomisation may still leave groups unbalanced on that variable by chance. This section introduces two formal designs: one that randomises freely, and one that first groups subjects to guarantee balance on a key variable before randomising.

Definition 9.23 *Completely Randomized Design (CRD)*
All subjects are randomly assigned to treatments **without** grouping first. Use when subjects are relatively similar and no known variable strongly affects the response.

When subjects are homogeneous, randomisation alone can be trusted to create balanced groups. Here are three situations where CRD is appropriate:

- A researcher tests two fonts on reading speed. She recruits 60 university students with similar reading levels. Since no strong confound is known, she randomly assigns 30 to Font A and 30 to Font B.
- Forty identical seedlings from the same seed batch are randomly assigned to two watering schedules. Because the plants are genetically uniform and grown in the same soil, no grouping is needed before assignment.
- A company tests two email subject lines on 500 subscribers drawn at random from the same mailing list. Subscribers are similar in engagement history, so treatments are assigned at random without any prior grouping.

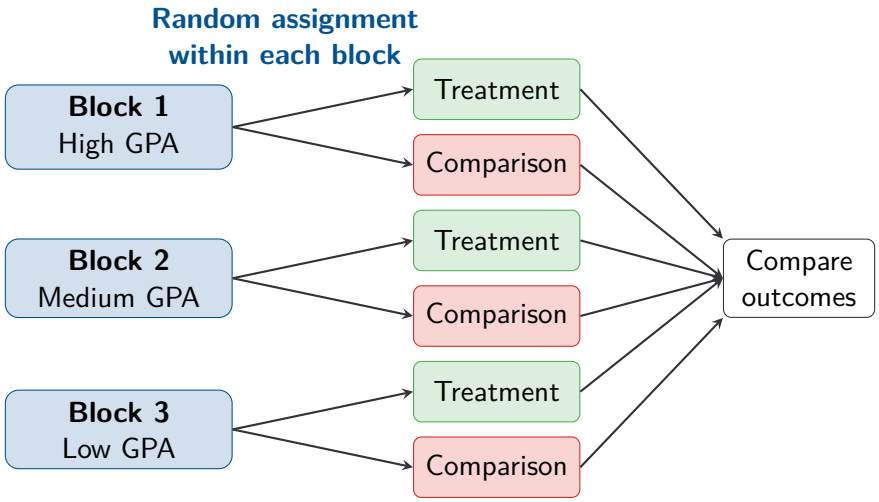
CRD relies on randomisation to balance groups. But when subjects vary on a characteristic that is known to drive the response, that variability adds noise to the comparison and makes a real treatment effect harder to detect. Blocking solves this: sort subjects into similar groups *first*, then randomise *within* each group, so every comparison is between units that started in the same place.

Definition 9.24 *Randomized Block Design (RBD)*
Subjects are first **grouped into blocks** of similar units, then randomly assigned to treatments **within** each block. Use when a known variable affects the response and you want to remove its influence from the comparison.

The blocking variable is what drives the grouping. Any variable known to affect the response is a candidate:

- A researcher tests a tutoring app on 60 students. Since prior GPA is known to affect academic performance, she groups students into high, medium, and low GPA blocks, then randomly assigns app vs. no app within each block.
- An agricultural trial tests two irrigation methods across 60 wheat plots spread across three soil zones (sandy, loam, clay). Since soil type affects water retention and yield, plots are blocked by soil zone before being randomly assigned to an irrigation method.
- A clinical trial tests a pain medication on patients aged 25–75. Age affects baseline pain sensitivity, so patients are grouped into three age blocks (25–40, 41–60, 61–75) and randomly assigned to treatment within each block.

The diagram below shows the RBD structure for the GPA example: randomisation happens *within* each block, not across the whole group.



Within each block, both groups share the same background on the blocking variable. Any difference in outcomes cannot be explained by it. Blocking removes a known source of variability from the comparison.

Key point: Block when a known variable is related to the response; use a CRD when subjects are relatively homogeneous and no such variable is known.

Remark
Blocking helps most when the blocking variable is *strongly* related to the response. If it is not, blocking adds complexity without much benefit.

9.6.1 Matched Pairs Design

The tightest possible block is a pair of just two units, matched as closely as possible or even the same unit measured twice. When every block has exactly two observations, we call it a matched pairs design.

Definition 9.25 *Matched Pairs Design*
A **matched pairs design** compares two treatments using natural pairs. It takes two forms:

1. **Paired subjects:** Two similar units are matched into a pair, and one is randomly assigned to each treatment.
2. **Within-subject (crossover):** Each unit receives *both* treatments in random order, and the two responses are compared.

Key point: Matched pairs is a special case of a block design in which each block contains exactly two units.

1. **Paired subjects example:** Match two runners by fitness level, then randomly assign one to new shoes and the other to standard shoes.
2. **Within-subject example:** Each person tastes coffee A and coffee B in *random order*, with a palate cleanser between tastings to reduce carryover effects; compare ratings.

When each person serves as their own control (version 2), many person-to-person differences are removed, making treatment effects easier to detect.

Example 9.26 *Matched Pairs: Paired Subjects*
Context: Researchers want to test whether a new study app improves quiz scores. They recruit 30 pairs of students, matching each pair by prior GPA. Within each pair, one student is randomly assigned to use the app; the other uses standard study materials. Quiz scores are compared within each pair.

(a) What are the two treatments?

(b) What is the matching variable, and why do we use it?

Example 9.27 *Matched Pairs: Within-Subject Design*
Context: A researcher tests whether background music affects reading comprehension. Each of 25 participants completes two reading passages of similar difficulty: one with background music and one in silence. The order of conditions is randomly assigned for each participant to control for carryover effects. Comprehension scores are compared within each person.

(a) What are the two treatments?

(b) How many blocks are there, and what is in each block?

Example 9.28 *Fertilizer and Tomato Varieties*
Context: A researcher wants to test if a new fertilizer increases tomato yield. She has 60 plants: 20 Beefsteak, 20 Cherry, and 20 Roma. It is known that tomato variety affects yield.

(a) Should she use a CRD or a block design? Why?

(b) What are the blocks?

Example 9.29 *When to Block?*
Task: For each experiment, decide whether to use a CRD, block design, or matched pairs design.

1. Testing two teaching methods on exam scores. Students come from different year levels (1st, 2nd, 3rd year).

---

2. Each participant tastes two brands of orange juice in random order and rates the flavour.

---

3. Testing whether a brief mindfulness exercise improves focus. Participants are 50 students from the same second-year psychology course, all with similar prior academic records and no known variable that strongly affects the response.

Exercise 9.30 *Designing a Matched Pairs Study*
A shoe company claims their new running shoe reduces 5K race times. A researcher recruits 40 runners of varying ability levels.

(a) Explain how to set up this study as a matched pairs design.

(b) Why is matched pairs better than a CRD here?

9.7. Blinding

Even with comparison, randomization, replication, and blocking, one more threat remains: knowing which treatment you received can change how you behave. Consider a study testing whether caffeine improves reaction time. Participants who know they got caffeine might try harder on the task. Participants who know they got a placebo might not bother. Either way, the results reflect knowledge of the treatment, not the treatment itself.

Definition 9.31 *Placebo*
A **placebo** is an inactive substitute that looks and feels identical to the real treatment (e.g., a sugar pill or saline injection). It ensures that any observed difference between groups reflects the treatment itself, not subjects' expectations about receiving it.

A placebo makes it possible to conceal treatment assignment from participants. The broader strategy of keeping one or both parties unaware of who received which treatment is called blinding.

Definition 9.32 *Blinding, Single-Blind, and Double-Blind*
Blinding keeps subjects, researchers, or both unaware of which treatment a subject receives, preventing expectations from influencing results.

- **Single-blind:** Either the subjects *or* the researchers do not know which treatment was applied, but not both.
- **Double-blind: Neither** the subjects **nor** the researchers who evaluate outcomes know which treatment was applied.

Why Blinding Matters

Without blinding (subject bias)	Without blinding (researcher bias)
• Subjects who *know* they received the treatment may try harder (expectancy bias) or feel better simply because they believe the treatment works (placebo effect) • Subjects in the control group may feel discouraged or change their effort • Expectations change behaviour	• Researchers may unconsciously favour the treatment group • Outcomes may be assessed more favourably for treatment subjects • Encouragement may differ between groups

Blinding is not always possible:

- **Surgery vs. physical therapy** : patients know which treatment they received.
- **Online vs. in-person teaching**: students cannot be hidden from the format.

When blinding is impossible, researchers should acknowledge this as a limitation and, where feasible, blind the people *measuring* outcomes (single-blind on the researcher side).

Example 9.33 *Blinding a Caffeine Study*
Context: Researchers want to test whether caffeine improves reaction time on a computer task.

(a) A lab assistant prepares two identical-looking capsules and gives one to each participant. Participants do not know which capsule contains caffeine, but the researcher running the reaction-time test does. Is this study single-blind or double-blind?

_____

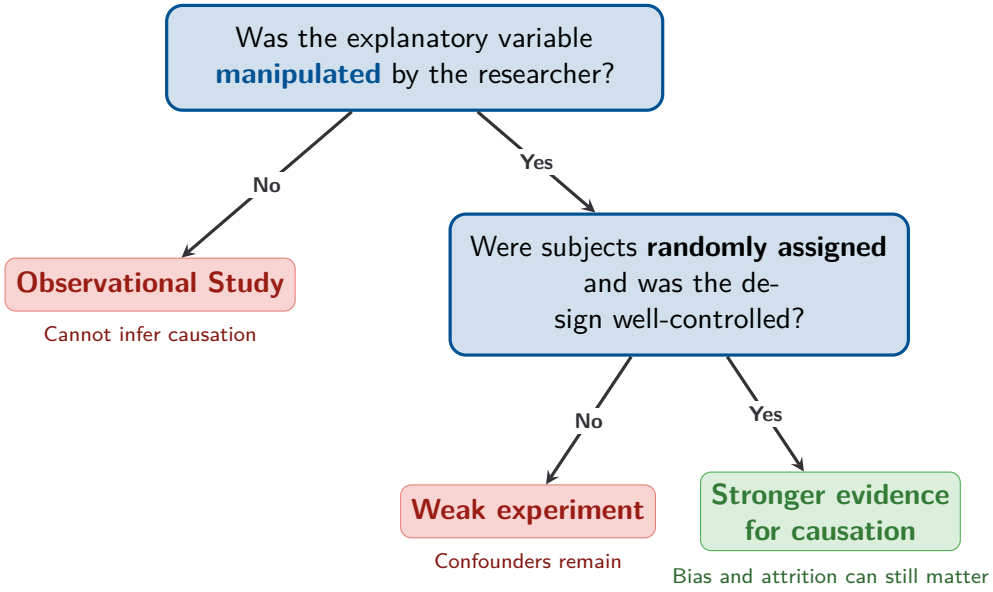
(b) In a revised version, the lab assistant uses coded labels so that neither the participants nor the researcher knows which capsule contains caffeine until all data have been collected. Is this study single-blind or double-blind?

_____

Key point: Blinding helps reduce bias caused by expectations. Double-blind studies provide the strongest protection against bias from both subjects and researchers.

9.8. Practice and Review

The flowchart below gives a quick way to evaluate a study design. When looking at any study, ask:



Note that a even experiments that are well designed might still have limitations. Drop-outs, protocol violations, or inconsistent measurement can weaken the evidence and hence the conclusions drawn.

Chapter 9: Key Concepts

- **Explanatory variable → Response variable:** The explanatory variable is the predictor; the response variable is the measured outcome.
- **Observational study vs. Experiment:**
  - Observational studies observe without manipulation → can show **association**
  - Experiments manipulate the explanatory variable → can provide evidence for **causation**
- **Confounding variable** (also called a **lurking variable**): Associated with both the explanatory and response variables; makes it very difficult to isolate the true cause of the observed effect.
- **Four Principles of Experimental Design:**
  1. **Comparison:** Use treatment and comparison groups.
  2. **Randomization:** Randomly assign subjects to balance confounders.
  3. **Blocking:** Group similar subjects before randomizing to control known sources of variation.
  4. **Replication:** Use enough subjects to detect a real effect.
- **Factors, Levels, and Treatments:** The number of treatments equals the product of the number of levels across all factors.
- **Design Types:**

  - **CRD:** Random assignment without grouping (subjects are similar).
  - **Block design:** Group by a known source of variation, then randomize within blocks.
  - **Matched pairs:** Compare two treatments using natural pairs (paired subjects or within-subject crossover).
- **Blinding:**
  - **Single-blind:** Subjects *or* researchers are unaware of assignments.
  - **Double-blind:** Neither subjects *nor* researchers know, which provides the strongest protection against bias.
  - A **placebo** is a fake treatment given to the control group.