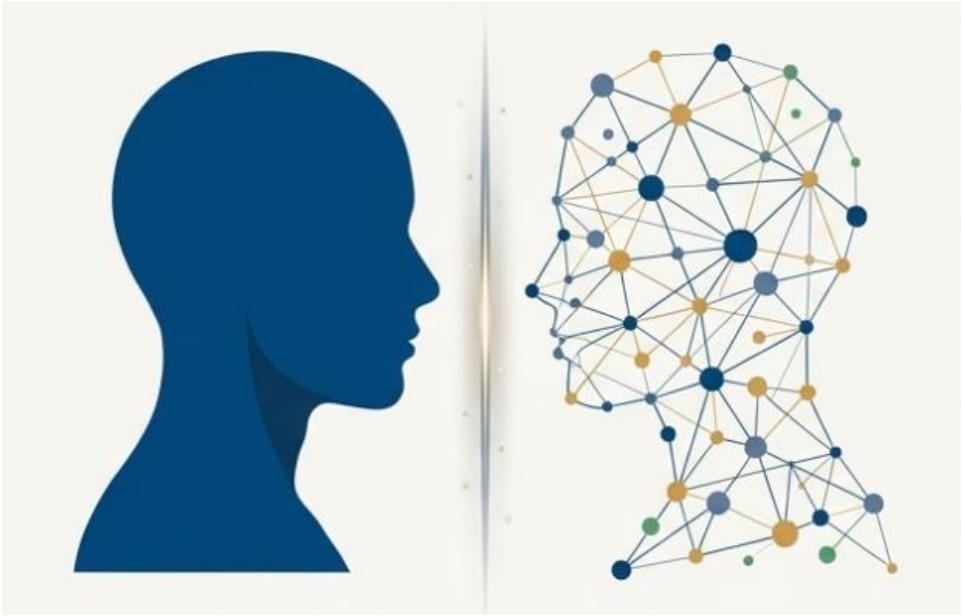
Treasury Board of Canada Secretariat | Secrétariat du Conseil du Trésor du Canada

Canada



# INTRODUCTION TO QUASI-EXPERIMENTAL RESEARCH DESIGN

*Measuring impact when randomization is not an option*

**Syeda Batool -** Expenditure Management Sector (TBS), **Ryan Kelly -** Strategic Policy Sector (ISED) - February 09, 2026

# BEYOND GOOD INTENTIONS: FROM *DID WE SPEND?* TO *DID IT WORK?*

Effective impact evaluation is critical. It helps policymakers decide whether programs are generating intended effects, promotes accountability, and fills gaps in our understanding of what works. The core goals are:

## Accountability

Proving to funders and the public that resources create real change, not just activity.

## Effectiveness

Learning what works and what doesn't to design better policies and programs in the future.

## Allocation

Making evidence-based decisions about scaling, modifying, or discontinuing programs.

# IMPACT EVALUATION

- Measures *causal effects*
- Focuses on **before vs. after** *and* **with vs. without**
- Applies a range of data and statistical methods

**Key Question:**

👉 *What changed because of the program?*

# IMPACT EVALUATION EXAMPLE

**Program: Government job training**
**Outcome: Earnings after 1 year**

| **Group** | **Average Earnings** |
|---|---|
| Participants | Higher |
| Non-participants | Lower |

**Impact = Difference in earnings**

Key Takeaway

📌 Impact ≠ correlation
📌 Impact = difference caused by the program

# IMPACT EVALUATION EXAMPLE

**Program: School textbook distribution program**

Goal: Improve student learning outcomes

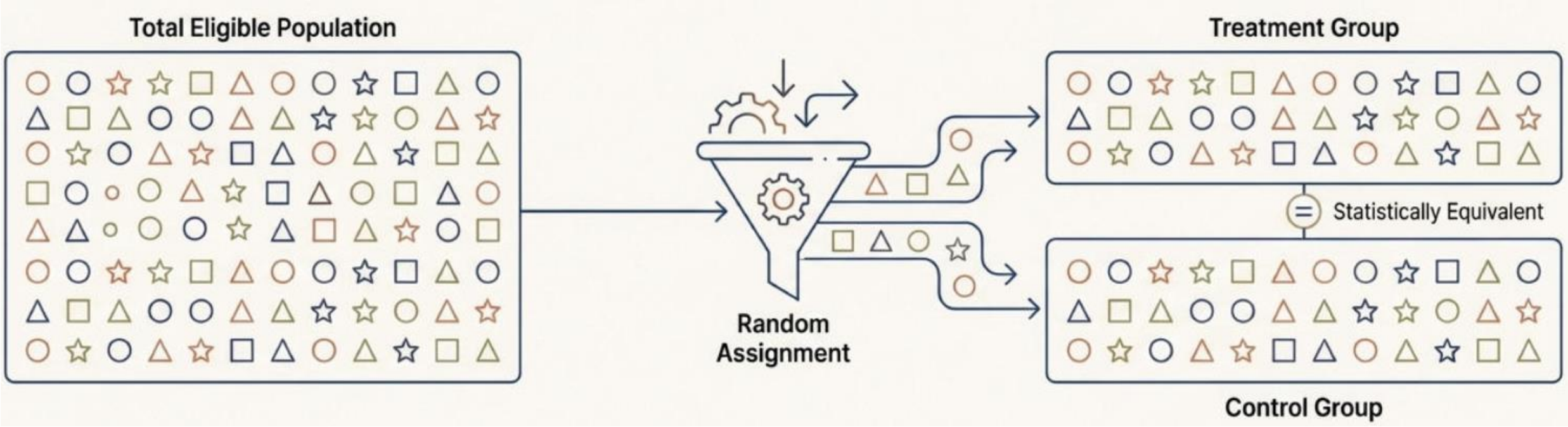
**Outcome of interest: Test scores at the end of the school year**

| Group | Received Textbooks? | Average Test Score |
|---|---|---|
| Treatment group | Yes | Higher |
| Control group | No | Lower |

**Impact = Difference in average test scores**

Key Evaluation Question 👉 *Did students learn more because of free textbooks, or would scores have improved anyway?*

# THE GOLD STANDARD: CREATING A PERFECT TWIN THROUGH RANDOMIZATION




Randomization ensures that assignment to treatment is unrelated to both observed and unobserved characteristics; in econometric terms, treatment assignment is **exogenous**.

# ALL THAT GLITTERS… PRACTICAL AND ETHICAL LIMITATIONS OF RANDOM ASSIGNMENT

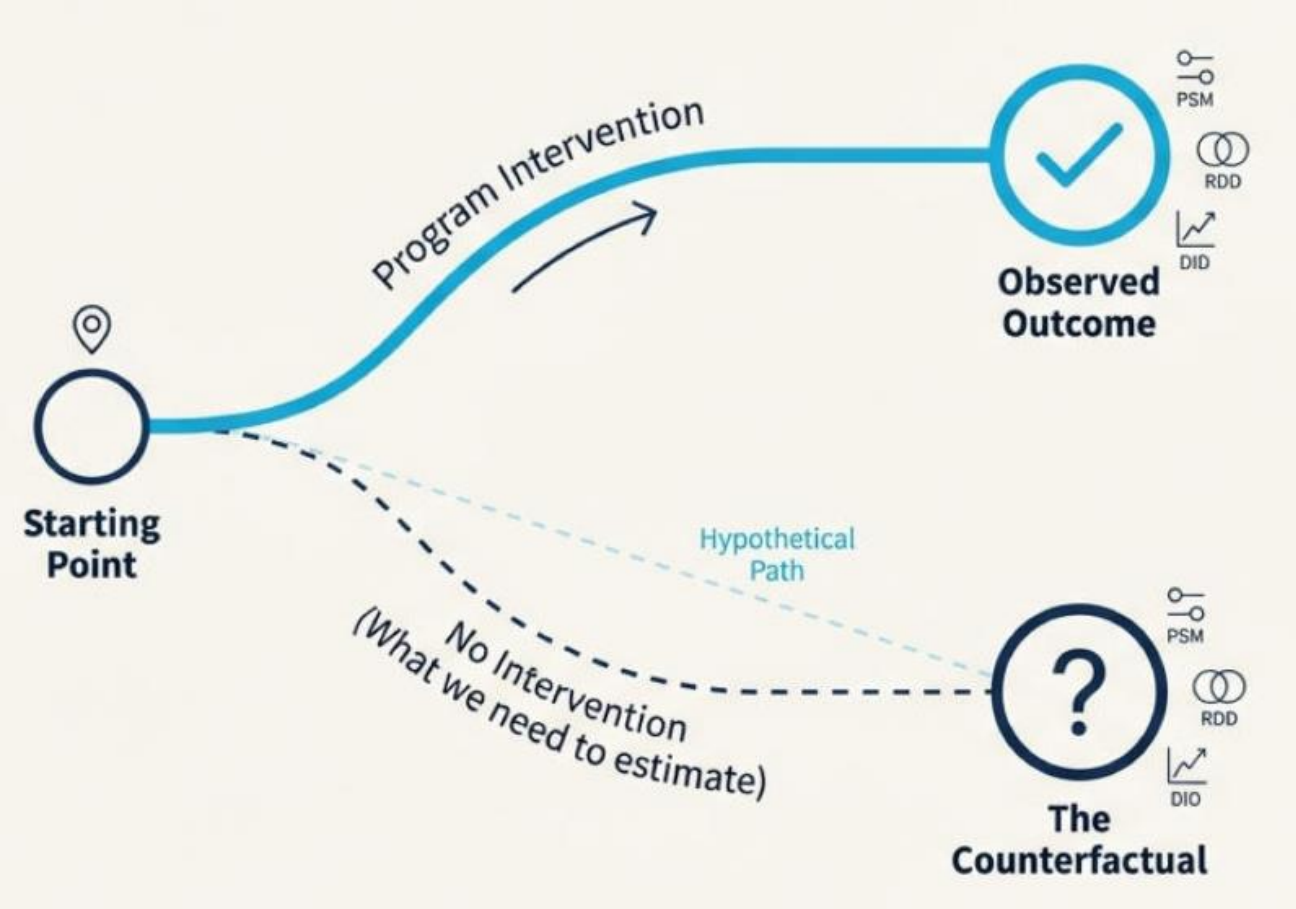
- Not always feasible or affordable
- Ethical and political constraints
- Limited scale or duration
- Implementation and compliance issues

# ANALYTICAL AND EXTERNAL VALIDITY: LIMITATIONS

- Limited external validity
- Average effects may mask heterogeneity
- Attrition and missing data
- Not always informative about mechanisms

# THE CORE CHALLENGE: THE UNSEEN WORLD OF THE COUNTERFACTUAL

***So, how do we confidently measure impact when we can't randomize?***



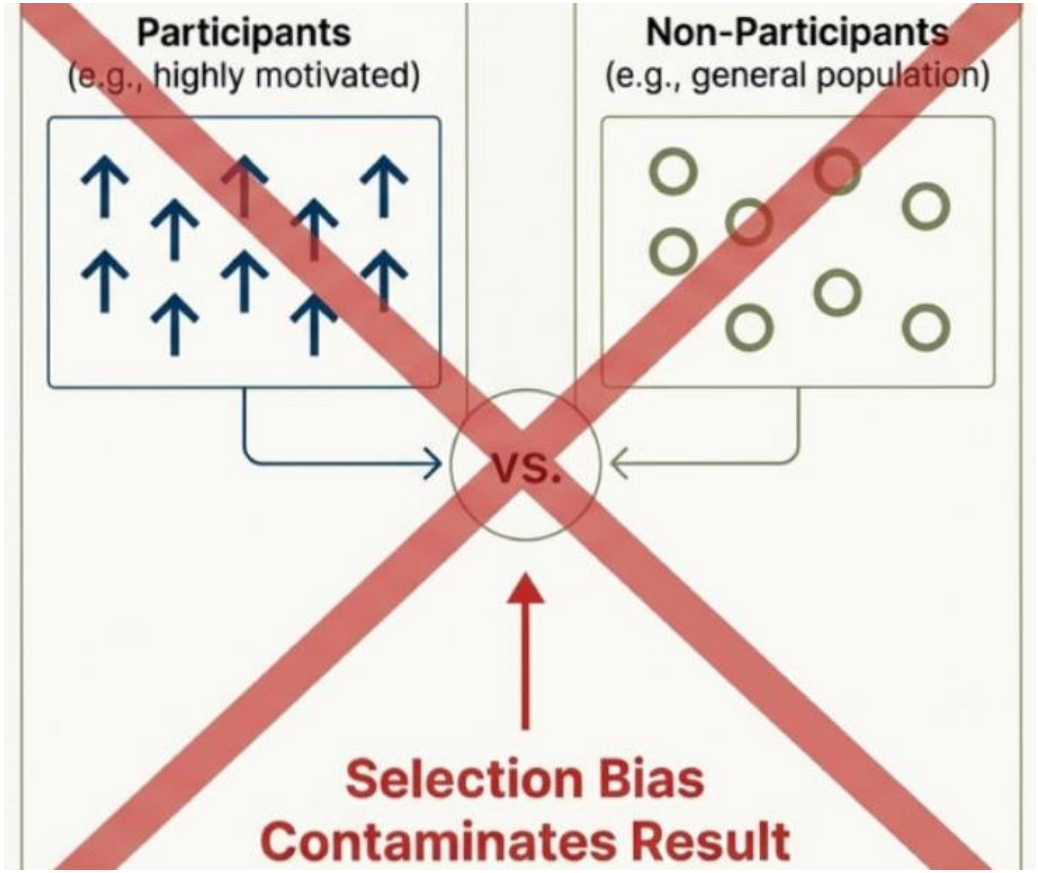

- Our goal in impact evaluation is to answer one question: *Did our program cause the observed change?*

- To do this, we must estimate the **counterfactual—** what would have happened without our intervention.

# WHY SIMPLE COMPARISONS FAIL: THE PROBLEM OF SELECTION BIAS

Program participants are often fundamentally different from non-participants *even before the program starts*. This systematic difference is **selection bias.**

👉 **Key idea:** Differences in outcomes may reflect *who joined*, not *what the program did*.



# WHY SIMPLE COMPARISONS FAIL: THE PROBLEM OF SELECTION BIAS

- **Self-Selection:** More motivated individuals might sign up for a training program.

- **Program Placement:** A microcredit program might be deliberately placed in the poorest villages.

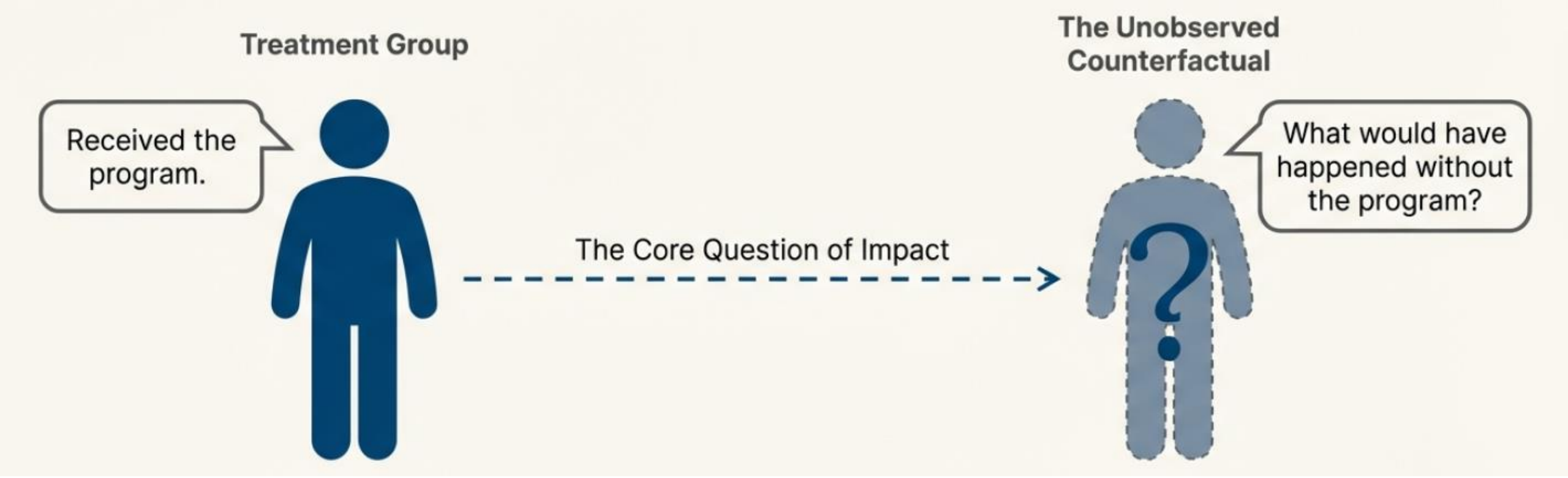
# WHY SIMPLE COMPARISONS FAIL: THE PROBLEM OF SELECTION BIAS

**Key Insight:** This bias contaminates simple comparisons, leading to wrong conclusions.

- **With-and-Without:** Compares participants to a non-equivalent group, potentially under- or over¬ estimating impact.

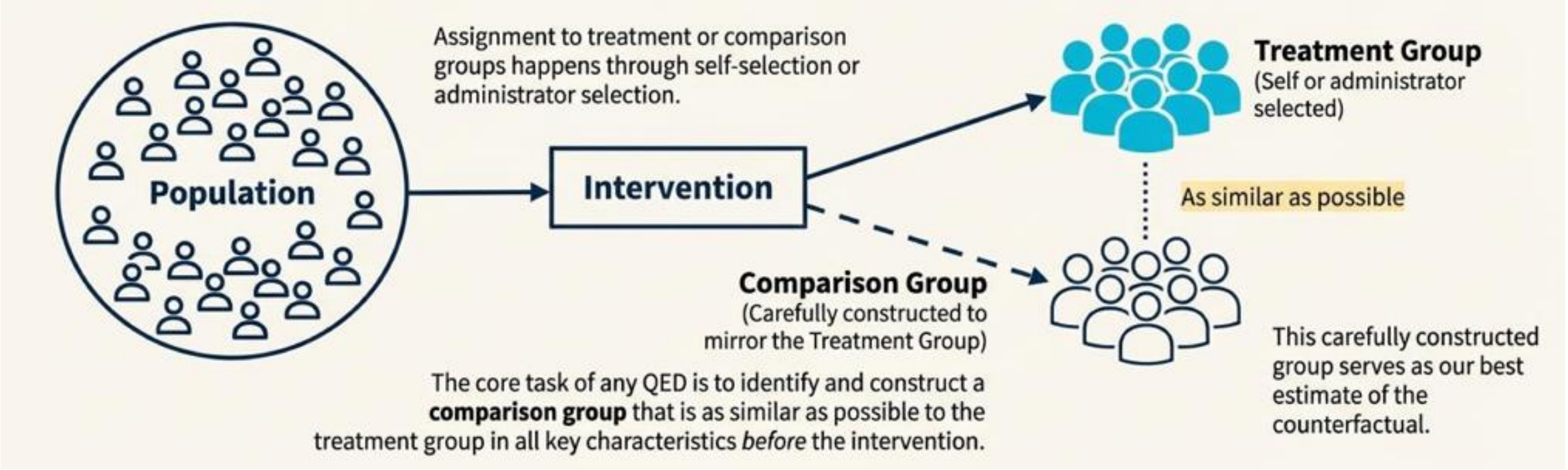
- **Before-and-After:** Fails to distinguish program effects from other external factors changing over time (e.g., economic recovery, other macro trends).

# THE COUNTERFACTUAL CHALLENGE




Impact evaluation is the science of creating a convincing comparison group to estimate the counterfactual!

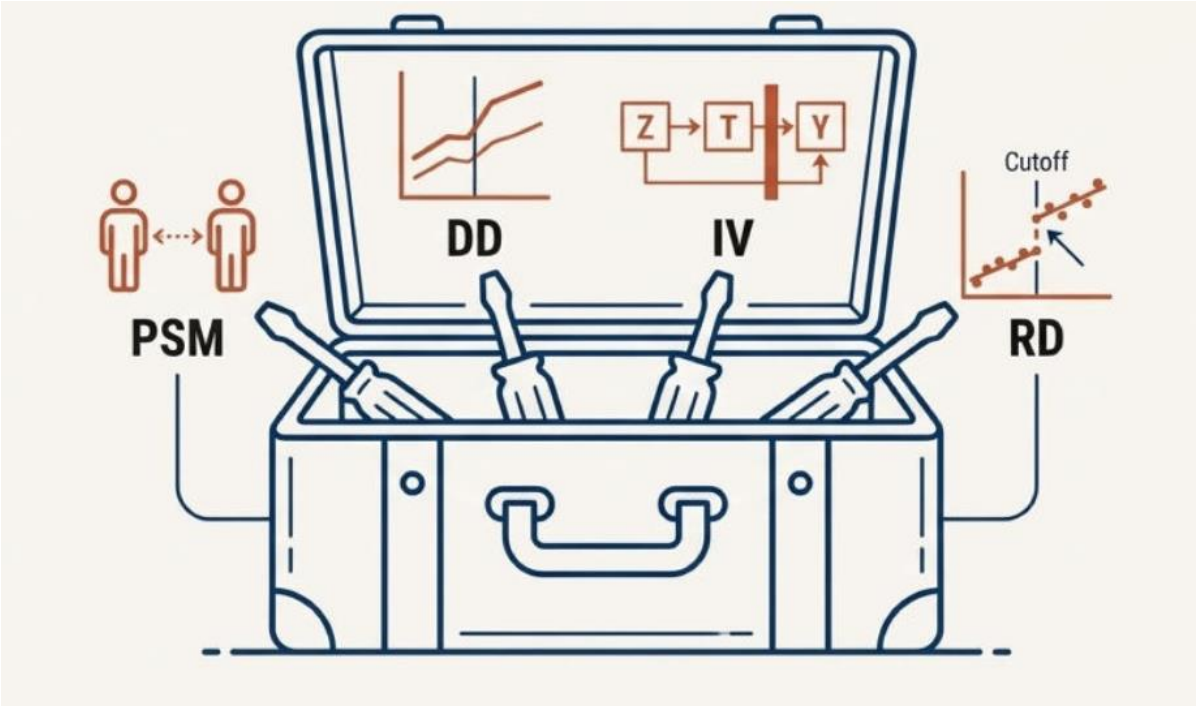
# THE SOLUTION: QUASI-EXPERIMENTAL DESIGN (QED)



# WHEN RANDOMIZATION ISN'T POSSIBLE: THE EVALUATOR'S TOOLKIT

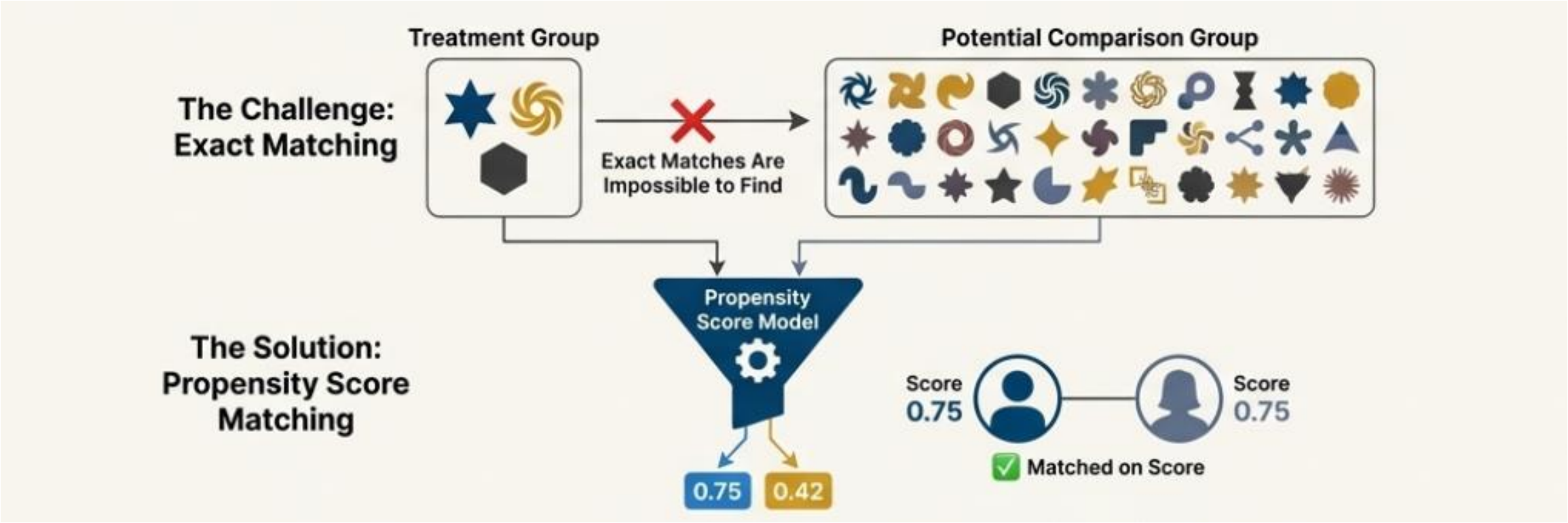
The following methods are powerful strategies to statistically 'mimic' a randomized trial by controlling for selection bias.

**Crucial Point:** Each method relies on different data and, most importantly, **different assumptions** about the nature of the selection bias. Choosing the right tool requires understanding these assumptions.



# METHOD 1: PROPENSITY SCORE MATCHING

# METHOD 1: PROPENSITY SCORE MATCHING




It’s like creating a "statistical twin" not based on an exact DNA match, but on a vast profile of observable traits.

# HOW PSM WORKS: THE 5-STEP PROCESS


**Survey & Sample**
Collect data on a large group of participants and non-participants.

**Estimate Scores**
Use a statistical model (logit/probit) to calculate a single "propensity score" for every individual their probability of participating given their observable traits.

**Match & Trim**
Match each treated person with one or more untreated people who have a very similar score. Discard anyone without a good match (this is called finding the "region of common support").

**Check Balance**
After matching, verify that the new treatment and comparison groups are now balanced (statistically similar) on their observable characteristics.

**Estimate Impact**
With the groups now balanced, the remaining difference in outcomes can be more confidently attributed to the program.

# CASE STUDY: THE INTERNSHIP ABROAD PUZZLE

**The Program:**
An intervention offers unemployed people a voluntary internship or short-term job contract abroad (IJA).

**The Goal:** To determine if the program had a real effect on participants' employability after they returned.

**The Challenge:** Participation is voluntary. This means participants may **self-select** into the program. We have data on 128 participants (the 'treated' group) and 272 non-participants (the 'control' group).

- Our outcome of interest is whether an individual is employed later.

# A FLAWED FIRST LOOK REVEALS A MAJOR PROBLEM

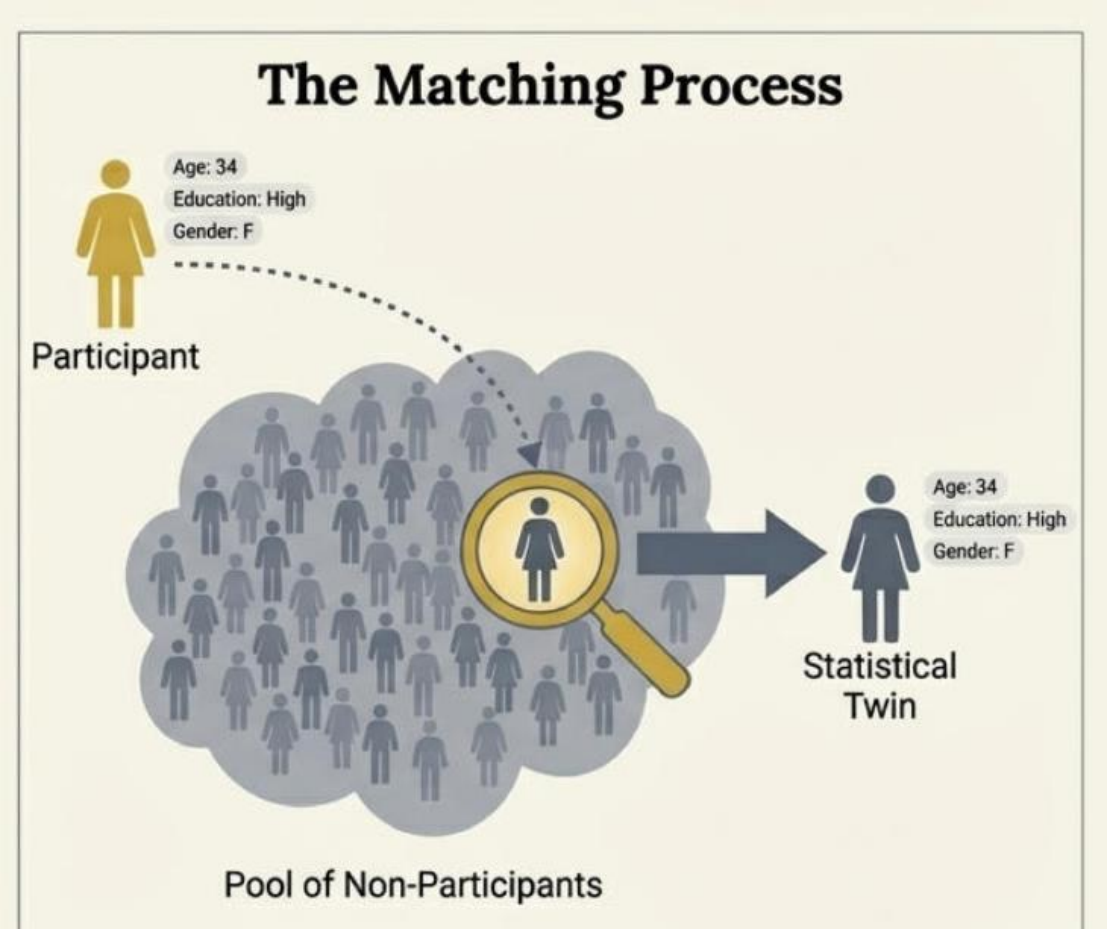
**Table: Mean Age of Participants vs Non-participants**

| Group | Mean Age | N |
|---|---|---|
| Participants (D=1) | 34.4 years | 128 |
| Non-Participants (D=0) | 43.4 years | 272 |
| **Difference** | **-9.04*** | |

(** indicates statistical significance)*


The Implication: We can't trust a simple comparison of employment outcomes. Are participants more employable because of the program, or simply because they are younger? This is selection bias in action.

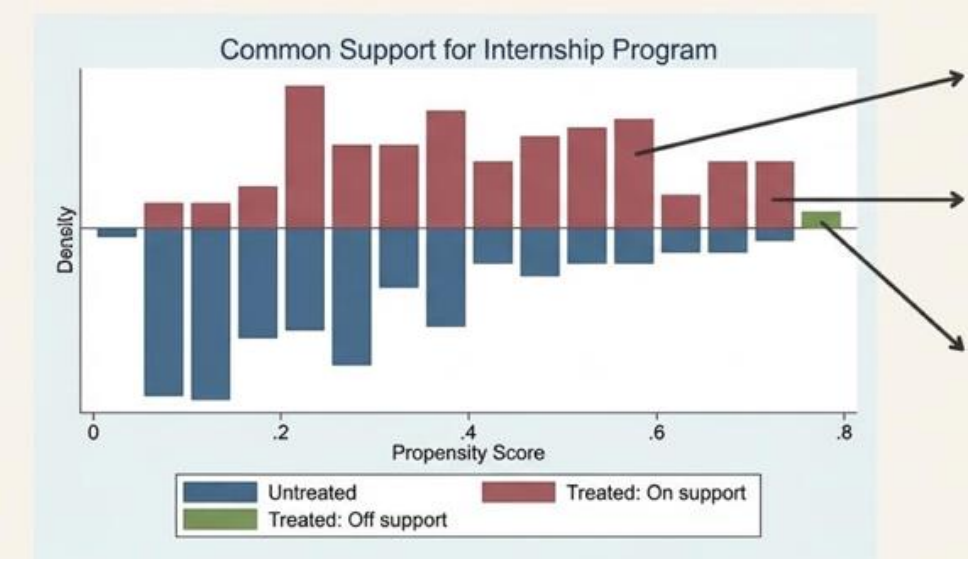
# SOLUTION: FINDING A "STATISTICAL TWIN" WITH PROPENSITY SCORE MATCHING



# PSM IN ACTION: THE STEPS TO A FAIR COMPARISON

Step 1: Estimate the Propensity Score for all individuals in the sample.

Step 2: Check the "Common Support" Condition




Blue bars: Distribution of propensity scores for untreated group.

Red bars: Treated individuals "On support" (good matches).

Green bar: Treated individuals "Off support" (too different, dropped from analysis).

Step 3: Choose a Matching Algorithm (e.g., Nearest Neighbors, Radius Matching) to pair participants with their "twins".

# THE MOMENT OF TRUTH: DID THE MATCHING CORRECT THE BIAS?

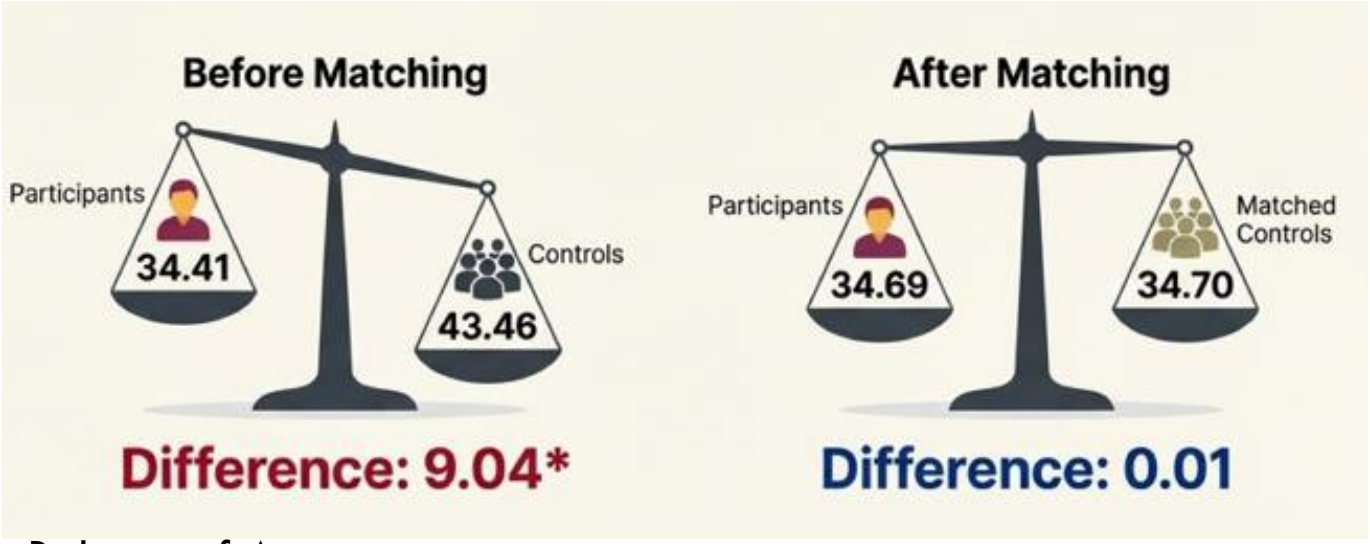



Table: Covariate Balance of Age

| | Participants (D=1) | Controls (D=0) | Difference |
|---|---|---|---|
| Before Matching | 34.41 | 43.46 | 9.04* |
| After Matching | 34.69 | 34.70 | 0.01 |

**The Result:** Success. The 9-year age gap has been **eliminated.** The propensity score has successfully balanced the two groups on this key characteristic. We now have a credible comparison group.

# THE VERDICT: THE INTERNSHIP PROGRAM HAD A SIGNIFICANT POSITIVE IMPACT

| **Table: Estimated Average Treatment Effect on the Treated (ATT)** | | |
|---|---|---|
| Matching Algorithm | Options | Impact (ATT) on Employability |
| Nearest Neighbors | 2 Controls | **+16.7 pp** ** |
| Nearest Neighbors | 4 Controls | **+21.8 pp** *** |
| Radius Matching | 0.02 radius | **+24.3 pp** *** |
| Kernel Matching | Function 2 | **+25.2 pp** *** |

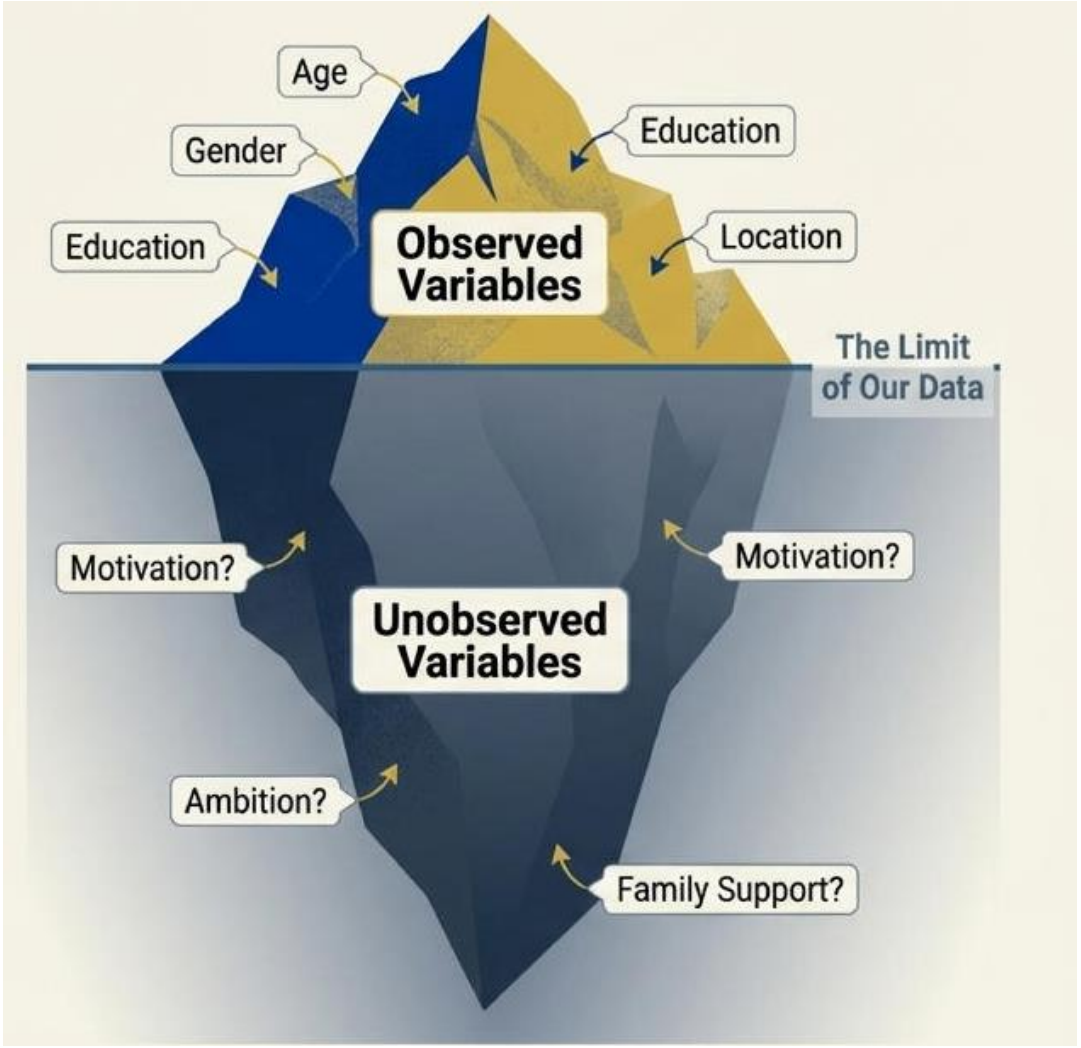
Note: pp = percentage points. ** and *** indicate significance.


**Conclusion:** The Internship Job Abroad (IJA) program increased the probability of being employed by between 17 and 25 percentage points for participants

# THE BIG "IF": PSM'S MOST IMPORTANT ASSUMPTION

PSM is powerful, but its validity rests on one strong, untestable assumption:

The Conditional Independence Assumption (CIA) or "Selection on Observables"



# METHOD 2
# DIFFERENCE IN DIFFERENCE

# THE LOGIC OF DiD: USING A CONTROL GROUP TO NET OUT TIME TRENDS

DiD removes biases from comparisons between a treatment and control group by accounting for trends that affect both groups over time.

**How it Works:**

1.Calculate the change in the outcome for the treatment group before and after the intervention.

2.Calculate the change in the outcome for the control group over the same time period.

3. The treatment effect is the difference between these two differences.

**The "Double Difference" Calculation:**

**Effect = (Treated After – Treated Before) – (Control After – Control Before)**

# THE LOGIC OF DID: USING A CONTROL GROUP TO NET OUT TIME TRENDS

**Data Requirements:**

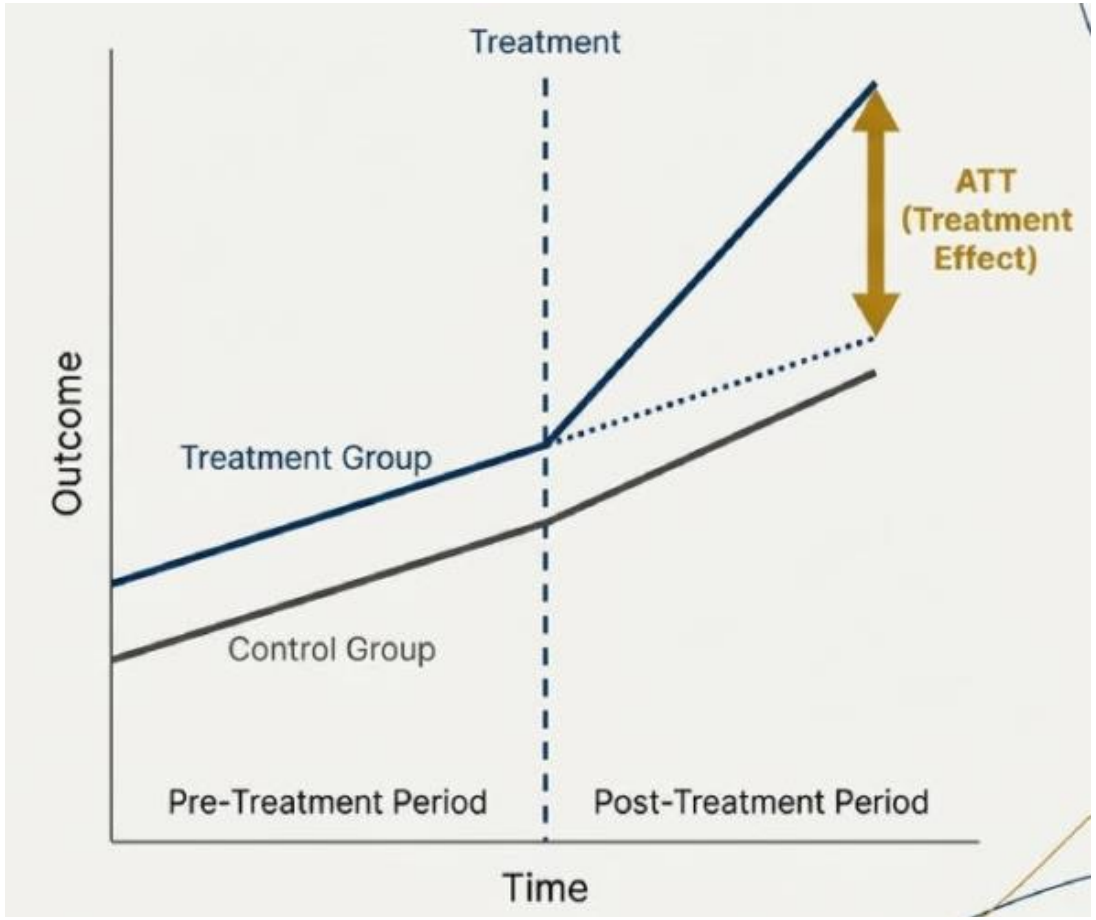
- Data must cover
  - before and after the intervention
  - Applies to both treatment and control groups

- Data Structure Options
  - Panel Data: Same individuals observed over time
  - Repeated Cross-Sections: Different individuals from the same groups over time

# THE CRUCIAL ASSUMPTION: PARALLEL TRENDS

**Key Assumption**

In the absence of the treatment, the average outcome for the treatment and control groups would have followed the same trend over time.



# CASE STUDY: THE ENTREPRENEURSHIP EDUCATION PARADOX

**•The Program:**

A leading entrepreneurship program (SMC) for college students in the Netherlands, designed to increase their skills and motivation for starting a business.

**Research Question:**

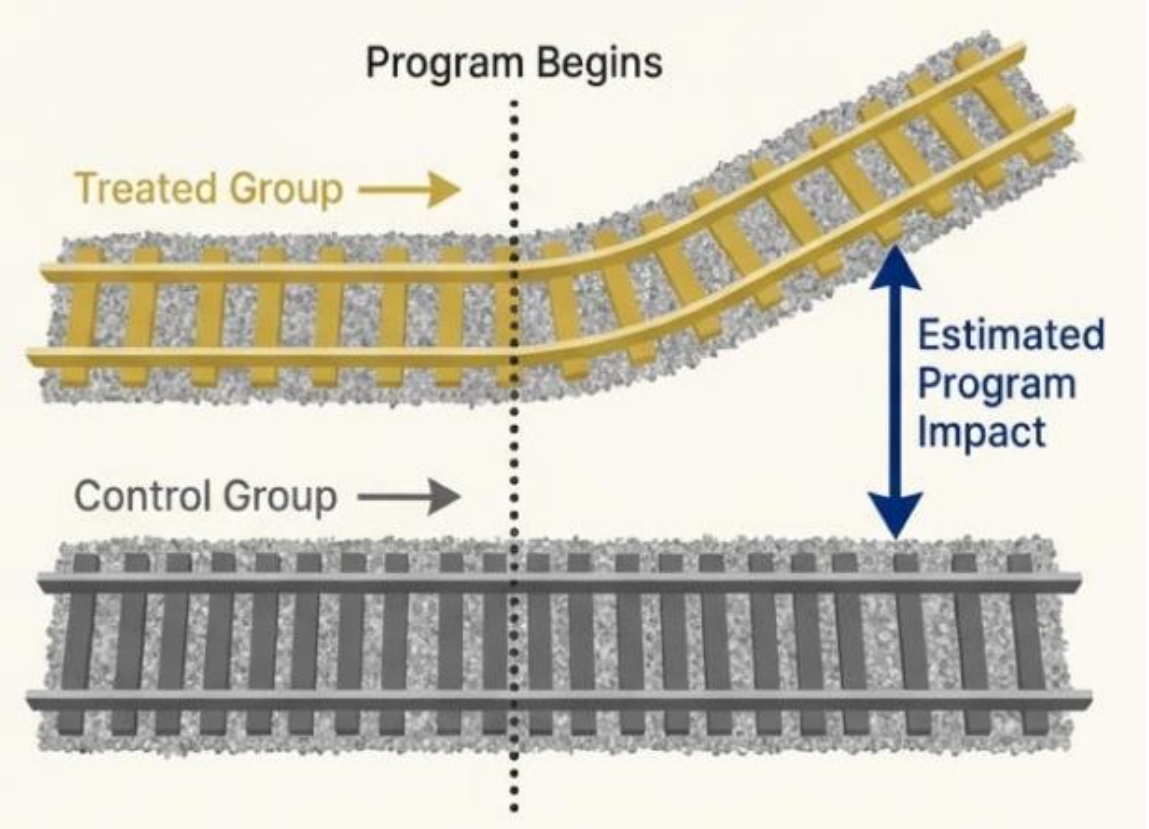
Does a student mini-company (SMC) program increase students' entrepreneurial skills and motivation?

The Setup:

- Treatment Group: Students at a Dutch vocational college campus in Breda, where the SMC program was mandatory.
- Control Group: Students at a different campus (Den Bosch) of the same college, which did not yet offer the program.
- Data: Surveys of students at both campuses were conducted before and after the academic year.

# SOLUTION: COMPARING TRAJECTORIES WITH DIFFERENCE-IN-DIFFERENCES

Instead of matching individuals, **DiD compares the change in the outcome** over time for the treated group to the **change in the outcome** for the control group.



# THE SURPRISING REVEAL: A NEGATIVE IMPACT ON INTENTIONS

The researchers compared the change in entrepreneurial skills and intentions between the Breda (Treated) and Den Bosch (Control) students. The results were not what policymakers expected.

While there were mixed results on specific skills, the most striking finding was on the primary goal:

| Outcome Measure | DiD Effect Estimate |
|---|---|
| Entrepreneur Skills | -0.188** |
| Creativity | -0.360** |
| Entrepreneurial Intentions | **-0.553*** |

The Finding: The program caused a statistically significant decrease in students' stated intentions to become entrepreneurs.

*(Source: Oosterbeek et al., 2010. Column 7 of Table 5)

# THE PARADOX EXPLAINED: A DOSE OF REALITY


- Program offered a realistic view of what it takes to run a business.
- Helped students realize entrepreneurship wasn't the right path for some.
- Important nuance: Effect stronger for female students; balancing business with other priorities led to lower entrepreneurial intentions.

The Takeaway: Rigorous evaluation doesn't just tell us if a program worked, it can help us understand how and for whom it worked.

# CASE STUDY: THE EFFECT OF MINIMUM WAGE ON EMPLOYMENT

The Classic Question (Card & Krueger,1994)

Does raising the minimum wage cause employment to fall?

The Natural Experiment

**Treatment:** In April 1992, New Jersey (NJ) increased its state minimum wage from $4.25to $5.05 per hour.

# CASE STUDY: THE EFFECT OF MINIMUM WAGE ON EMPLOYMENT

**Setup:** The researchers used this policy change to create a DiD study.

- **Treatment Group:** Fast-food restaurants in New Jersey.
- **Control Group:** Fast-food restaurants in eastern Pennsylvania (PA), where the minimum wage remained $4.25.

**Data:** They surveyed -400 fast-food stores in both states before (Feb 1992) and after (Nov 1992) the minimum wage increase.

# DID IN ACTION: UNPACKING THE DOUBLE DIFFERENCE

**Table: Average Full-Time Equivalent (FTE) Employment Per Store**

| | Pennsylvania (Control) | New Jersey (Treated) | Difference (NJ - PA) |
|---|---|---|---|
| Before (Feb 92) | 23.33 | 20.44 | -2.89 |
| After (Nov 92) | 21.17 | 21.03 | -0.14 |
| Change (After-Before) | -2.16 | +0.59 | +2.76 |

The First Difference (Control): Employment in PA (control) fell by 2.16 FTEs.

The First Difference (Treated): Employment in NJ (treated) ★rose*by 0.59 FTEs.

The Second Difference (The DiD Estimator): DiD Effect = (Change in NJ) - (Change in PA) = (+0.59) -(-2.16) = +2.76

Conclusion: Contrary to simple economic theory, the increase in the minimum wage in New Jersey did not decrease employment in fast-food restaurants; the evidence suggests it may have slightly increased it relative to the comparison group.

# THE STRATEGIST'S CHOICE: PROPENSITY SCORE MATCHING VS. DIFFERENCE-IN-DIFFERENCES

## Propensity Score Matching (PSM)

**Core Logic:** Selects a control group that looks the same on observed pre-treatment characteristics.

**Key Assumption:** Conditional Independence (Selection on Observables). Assumes no unobserved variables simultaneously affect participation and outcome.

**Data Needs:** Rich cross-sectional data with many pre-treatment covariates.

**Handles:** Selection bias based on measurable factors (age, education, etc.).

**Vulnerable to:** Unobserved characteristics like motivation, talent, family support.

## Difference-in-Differences (DiD)

**Core Logic:** Uses a control group to model what would have happened over time without treatment.

**Key Assumption:** Parallel Trends. Assumes treatment and control groups would have followed the same trend over time.

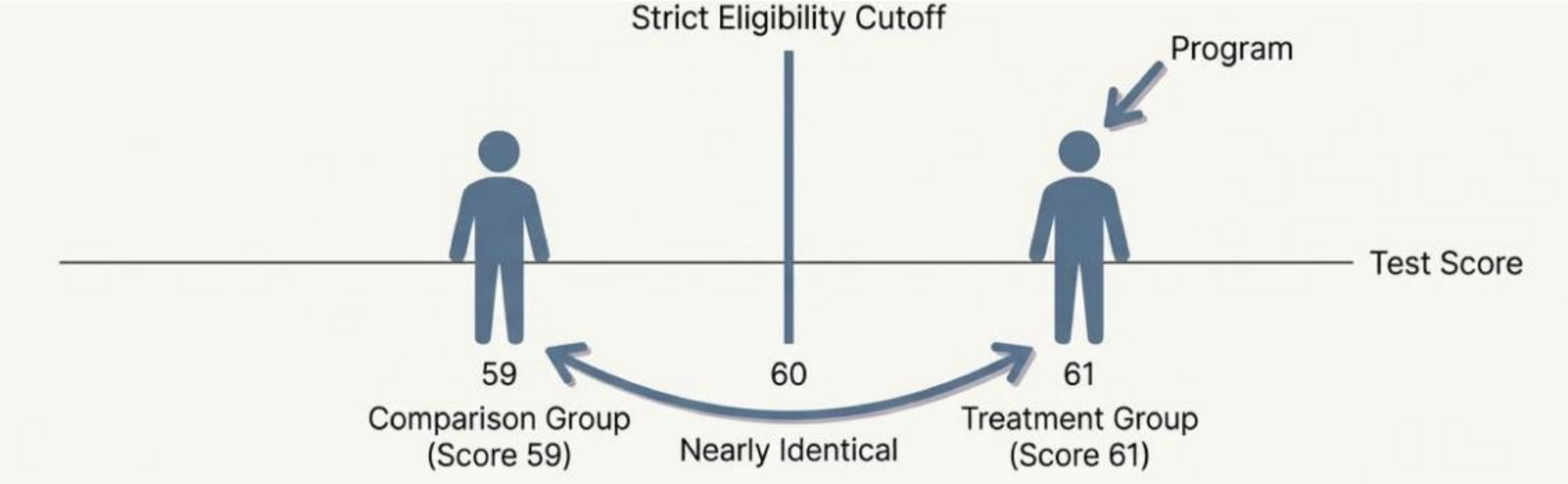
**Data Needs:** Panel data or repeated cross-sections (at least two time periods).

**Handles:** Selection bias from time-invariant unobserved characteristics (e.g., innate motivation).

**Vulnerable to:** Events or shocks other than treatment that affect only the treatment group over time.

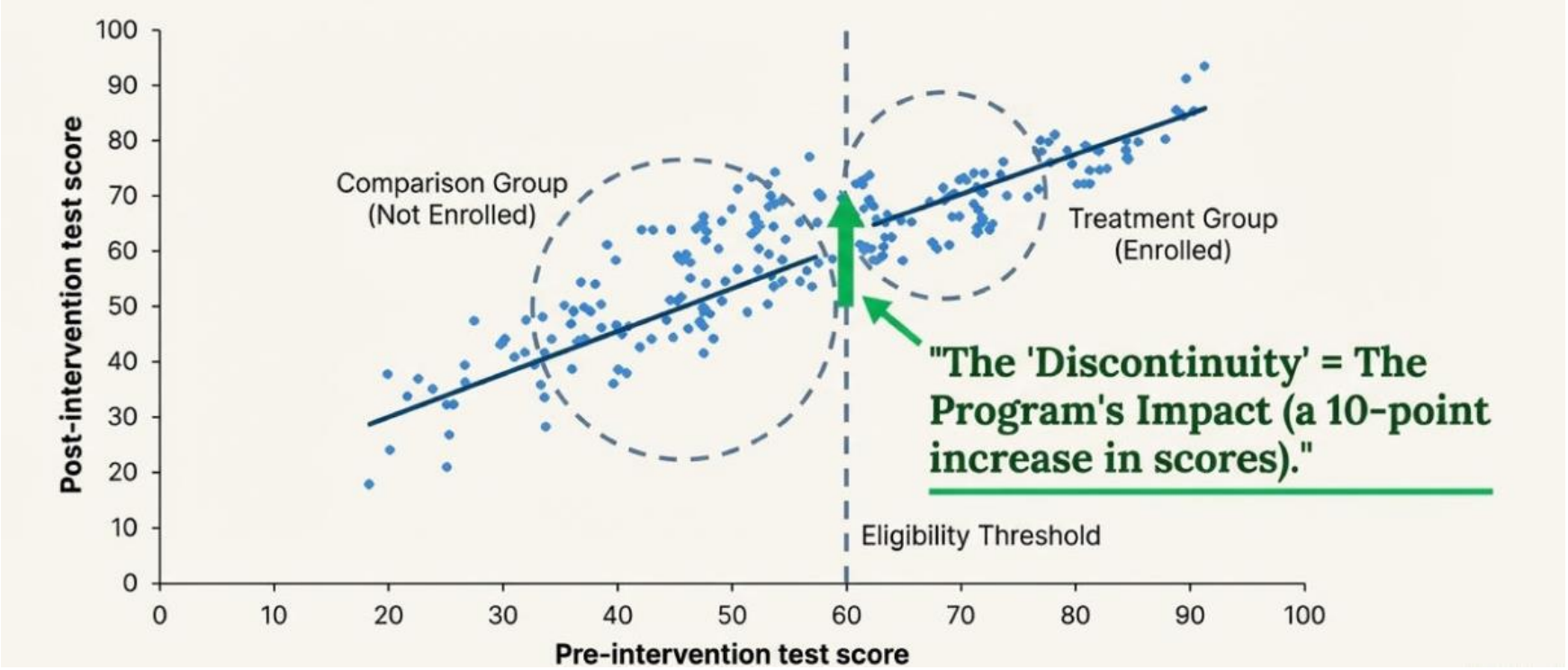
# METHOD 3: REGRESSION DISCONTINUITY DESIGN (RDD)

# METHOD 3: REGRESSION DISCONTINUITY DESIGN (RDD)



# RDD IN ACTION: THE POWER OF THE CUTOFF

A remedial education program is given to students scoring below 60 on a pre-test. We want to measure its impact on their post-test scores.



# ADVANTAGES AND DISADVANTAGES OF RDD

**Advantages**

**•Handles Unobservable**

Because people just around the cutoff are so similar, RDD is much better at dealing with unobserved characteristics than PSM. Its causal claims are often considered more credible.

**•Transparency**

The eligibility rule is clear and the analysis is visually intuitive.

**Disadvantages**

**• Local Effect**

The impact estimate is only valid for the population right around the cutoff. The effect might be different for those far from the threshold.

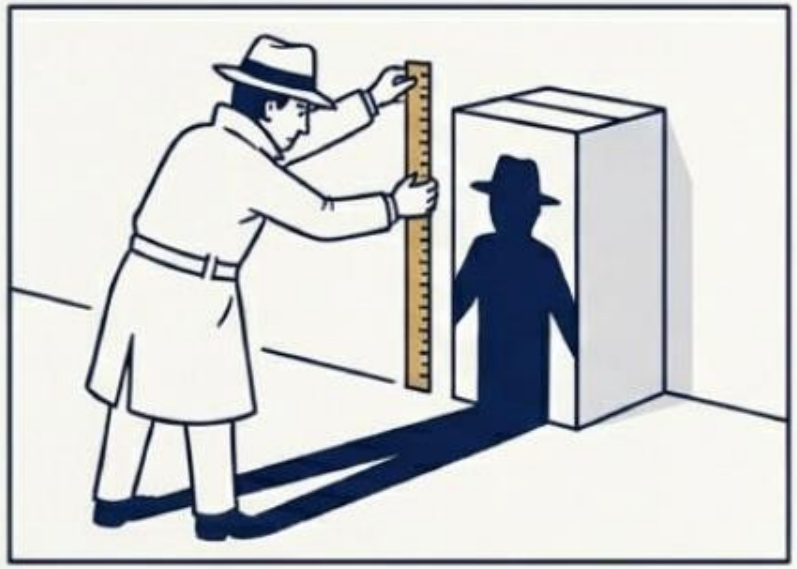
**• Limited Applicability**

Requires a program with a sharp, clearly-defined eligibility rule and enough data clustered around that cutoff.

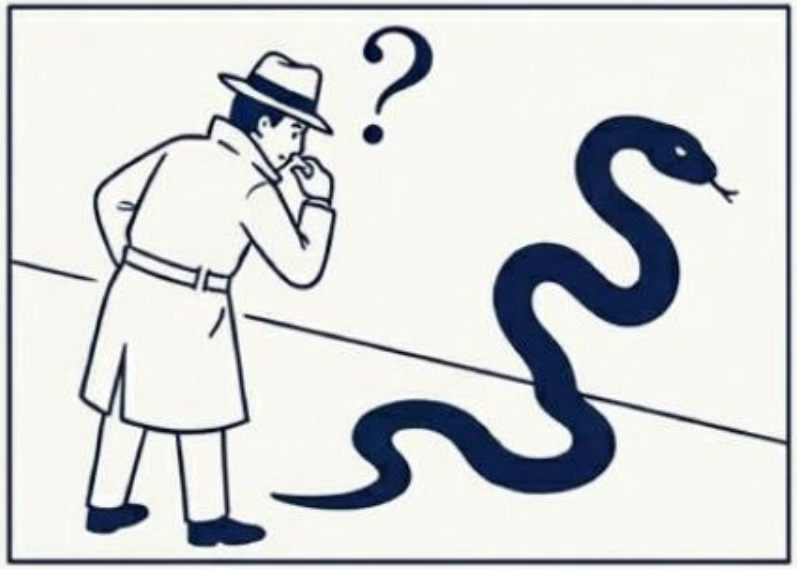
# THE TOUGHEST CHALLENGE: WHAT IF THE BIAS IS HIDDEN AND CHANGES OVER TIME?

- **PSM** fails if unobserved factors matter.
- **DiD** fails if the unobserved bias changes overtime (e.g., participants' motivation increases *because* of the program, or program placement targets areas with changing growth potential).

PSM: Finds what's under the light (Observables).



DiD: Measures a constant shadow (Time-Invariant Bias).

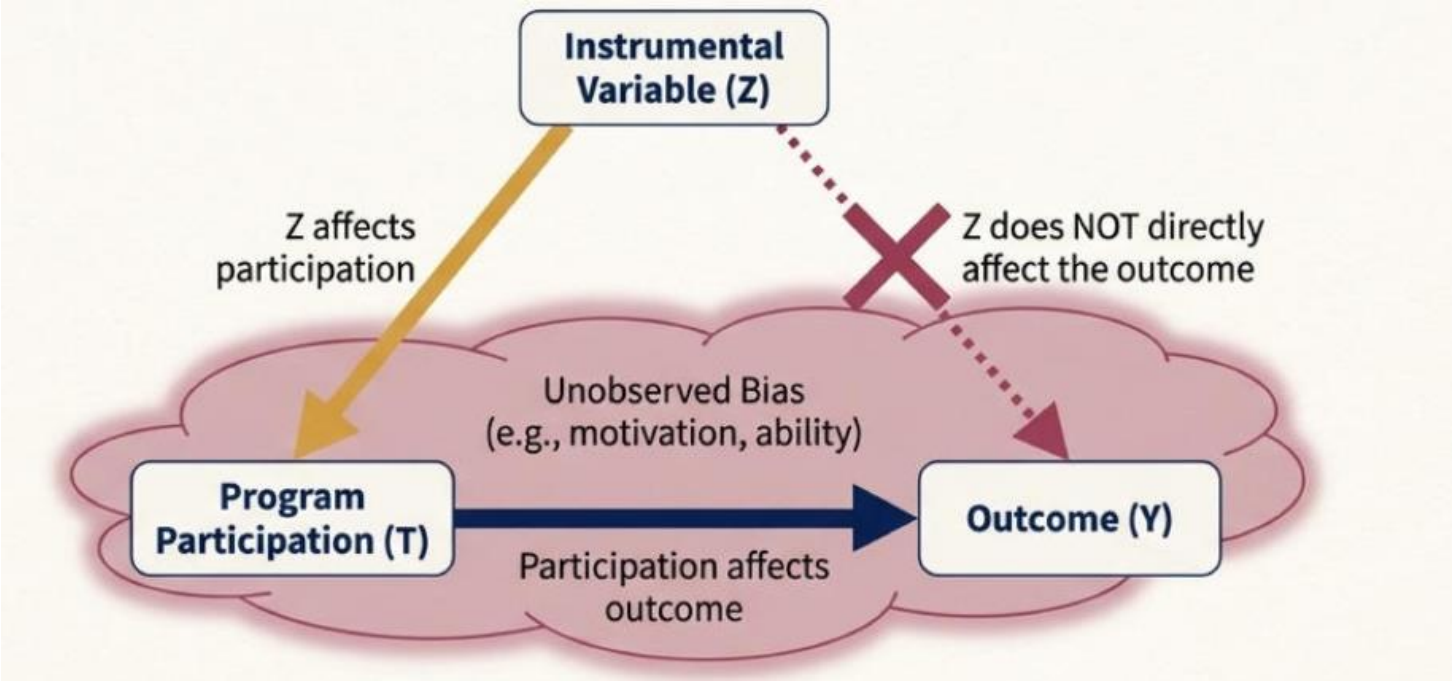


The Problem: A changing, hidden shadow (Time-Varying Bias).

# METHOD 4: INSTRUMENTAL VARIABLE (IV)

# THE SOLUTION: INSTRUMENTAL VARIABLES (IV)

**The Idea:** Find a source of variation—an “instrument"—that is as good as random. This instrument acts as a "random nudge/’ encouraging some people to participate but not others, without directly affecting their outcomes.

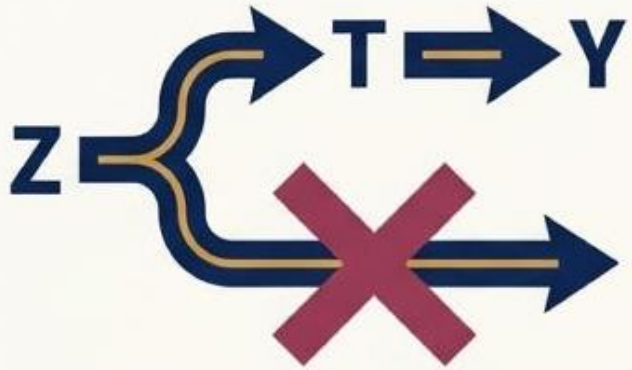


# THE TWO GOLDEN RULES OF A GOOD INSTRUMENT

To be a valid instrument, a valid variable must satisfy two strict conditions;


**1. The Relevance Condition**

$$\text{cov}(Z, T) \neq 0$$




**2. The Exclusion Restriction**

$$\text{cov}(Z, \varepsilon) = 0$$

# WHERE DO WE FIND THESE 'INSTRUMENTS' IN THE REAL WORLD?

Finding a valid instrument requires deep knowledge of the program's design and context. Common sources include:

## Geography of Program Placement

Sometimes, programs are rolled out in some regions but not others for reasons unrelated to the outcomes (e.g.,, administrative convenience).

Distance to a program center can be an instrument.

# WHERE DO WE FIND THESE ‘INSTRUMENTS’ IN THE REAL WORLD?

## Eligibility Rules

A program might have a sharp, arbitrary cutoff for eligibility (e.g., age, income, or land ownership).

Being just above or below the cutoff can serve as an instrument.

# WHERE DO WE FIND THESE ‘INSTRUMENTS’ IN THE REAL WORLD?

## Randomized Encouragement

Instead of randomizing the program itself, we can randomly give some people an incentive or extra information to encourage them to join.

The encouragement itself is the instrument.

# WHERE DO WE FIND THESE ‘INSTRUMENTS’ IN THE REAL WORLD?

## Policy Design

Features of a policy's implementation, like whether men and women must join separate groups in a microfinance program, can create exogenous variation in participation.

# A SPECIAL CASE: REGRESSION DISCONTINUITY (RD)

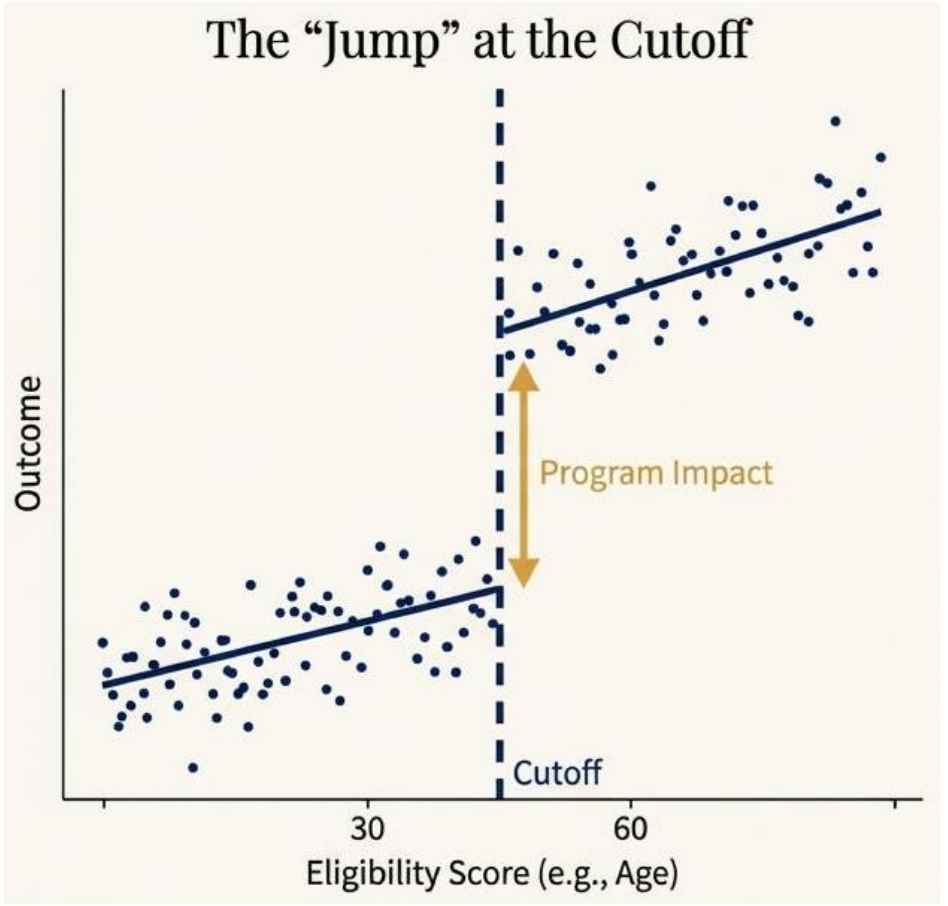
RD is a powerful design that uses an eligibility rule as an instrument.

**The Idea: We can compare people who are just barely eligible for a program with those who are just barely ineligible.**

**Example: Exploiting Eligibility Rules in South Africa**

A social pension program has a strict age cutoff for eligibility. It's reasonable to assume that people who are 60 years old (just eligible) are very similar to people who are 59 years and 11 months old (just ineligible) in all other respects. The difference in their outcomes can be attributed to the program.

In this case, the instrument is the eligibility cutoff itself. It powerfully predicts participation but is unlikely to be directly correlated with other factors affecting outcomes right around that cutoff point.

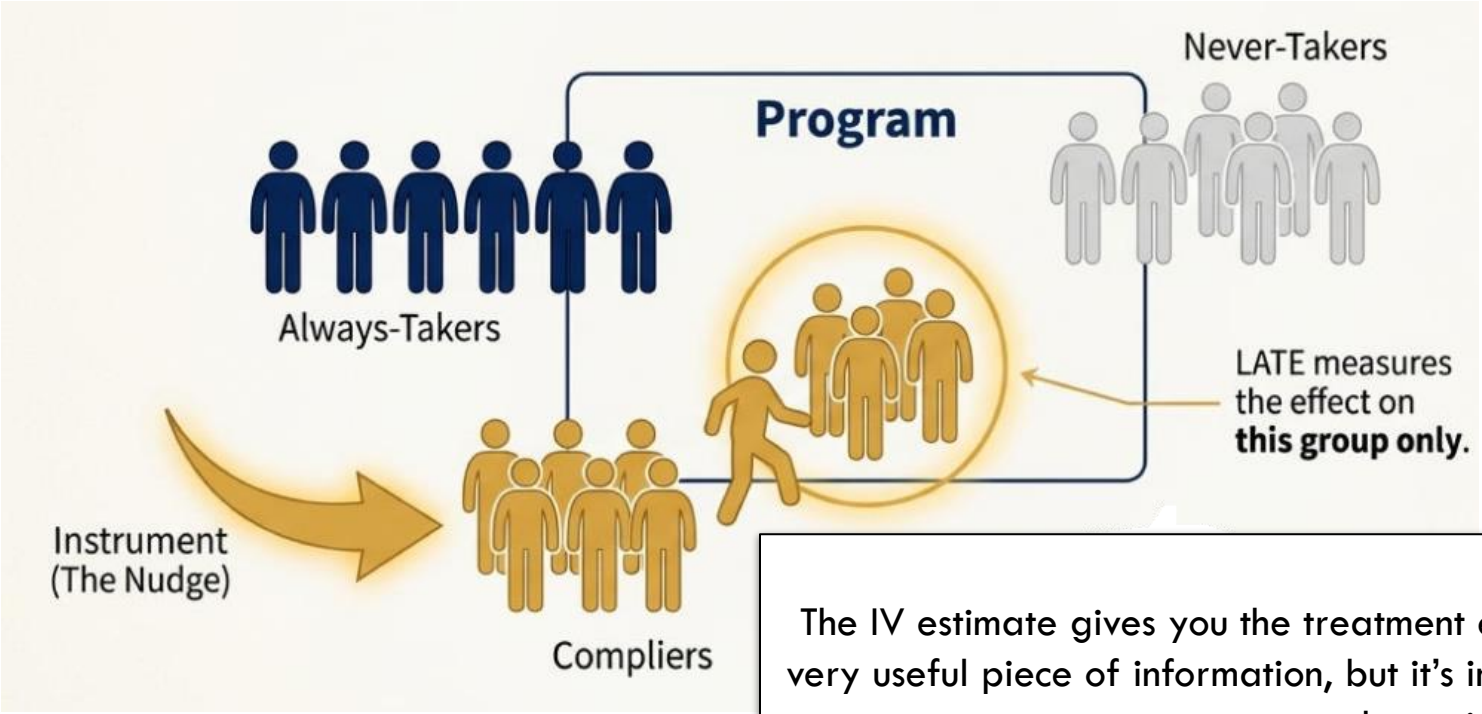


# A WORD OF CAUTION: THE DANGER OF "WEAK INSTRUMENTS"


It is better to have no instrument than a weak one. Researchers must rigorously test the strength of their proposed instruments before drawing conclusions.

# A CRUCIAL DETAIL: WHAT DOES IV ACTUALLY MEASURE?

An IV estimate does not measure the average effect for every participant. It measures the Local Average Treatment Effect (LATE): the average effect of the program for the specific subgroup of people who were induced to participate by the instrument.




The IV estimate gives you the treatment effect specifically for the "Compliers." This is a very useful piece of information, but it's important to remember it may not generalize to the entire population.

# THE SLEUTH'S TOOLKIT: A COMPARISON OF METHODS

| Method | How it Handles Selection Bias | Key Assumption | Typical Data Needs |
|---|---|---|---|
| Propensity Score Matching (PSM) | Controls for observable differences by creating a 'statistical twin' | Unconfoundedness: No selection on unobserved characteristics. | Rich cross-sectional data with many pre-program covariates. |
| Double Difference (DiD) | Controls for unobserved but time-invariant differences by comparing changes over time. | Parallel Trends: Treatment and control groups would have followed similar trends without the program. | Panel data or repeated cross-sections (data from before and after). |
| Instrumental Variables (IV) | Controls for unobserved and time-varying differences using an external source of variation. | Relevance & Exclusion: The instrument must affect participation but not the outcome directly. | Cross-sectional or panel data, plus a valid instrument. |
| Regression Discontinuity (RD) | A special case of IV that controls for bias by comparing units just above and below an eligibility cutoff. | Continuity: Units just on either side of the cutoff are comparable in all other respects. | Data on the 'running variable' that determines eligibility. |

# WHAT DATA DO YOU NEED?

**The quality of any impact evaluation hinges on the quality and type of data collected.**

**For all Non-Experimental Methods**

• **Rich Covariates:** Collect detailed data on household, individual, and community characteristics before the program begins. This is essential for PSM and for checking balance in other methods.

• **Common Survey** Instrument: Use the same questionnaire and survey methodology for both participant and nonparticipant groups to ensure comparability.

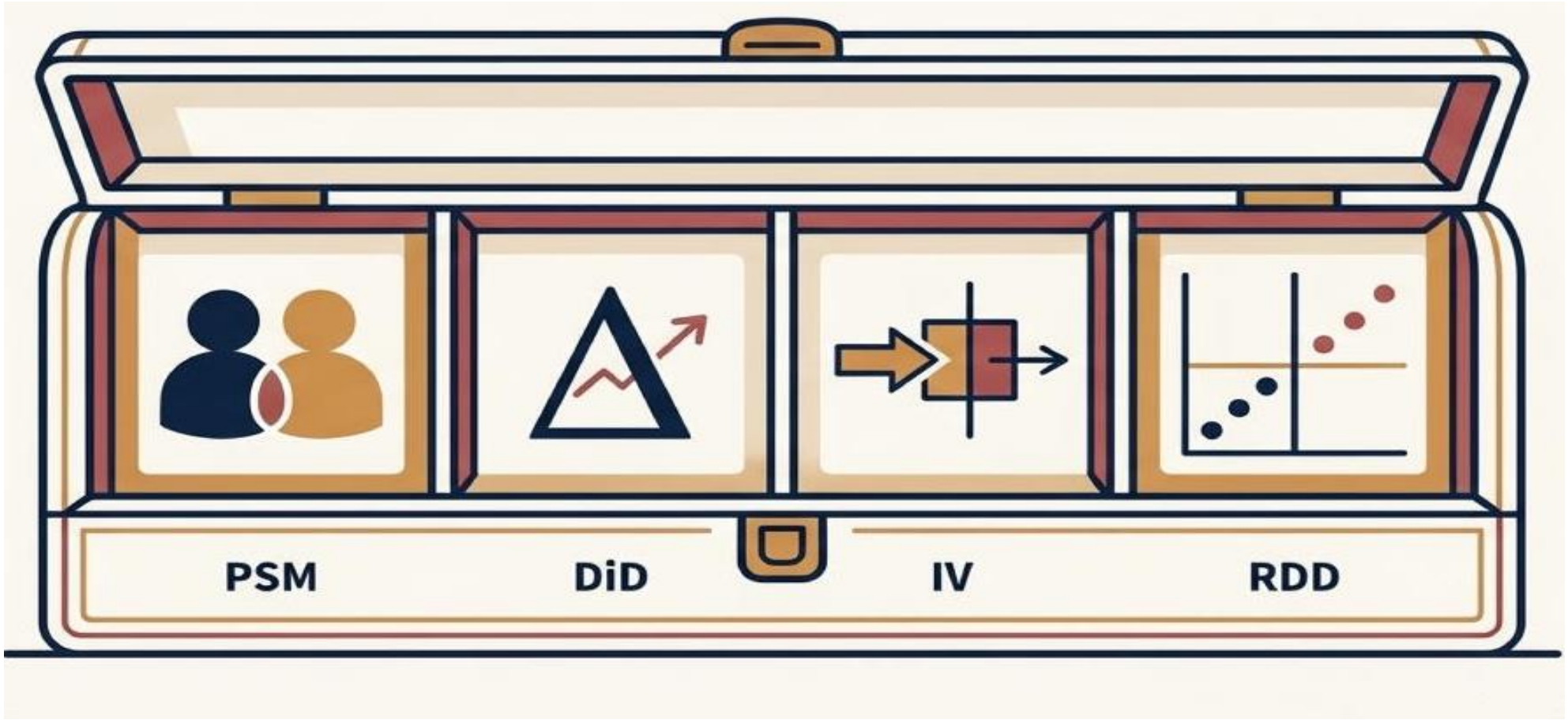
• **Large Sample** of Nonparticipants: A large, representative sample of eligible nonparticipants is crucial for finding good matches (PSM) and ensuring statistical power.

**For DiD and Panel IV**

• **A Baseline Survey** is Key: Collecting data before the intervention is critical. It allows you to test the parallel trends assumption (DiD) and control for initial conditions.

• **Panel Data:** Following the same individuals or households over time is the gold standard for controlling for unobserved heterogeneity.

# THERE IS NO SILVER BULLET, ONLY THE RIGHT TOOL FOR THE JOB



# THANK YOU

Syeda.Batool@tbs-sct.gc.ca
ryan.kelly@ised-isde.gc.ca