

Measuring What Matters: Impact Evaluation for Better Business Decisions

Appreciation Seminar

Melba V. Tutor, Resource Person
19 October 2026, PSRTI Quezon City

Learning Outcomes

At the end of the seminar, participants will be able to:

- Differentiate impact evaluation from routine monitoring, process assessment, and outcome tracking;
- Identify inputs, activities, outputs, and outcomes using a Theory of Change;
- Design basic evaluation questions, select key indicators, and outline the primary steps for conducting an impact evaluation;
- Appreciate counterfactuals, treatment vs. control groups, and common evaluation designs; and



Outline

- I. Introduction to Impact Evaluation
- II. Theory of Change: From Activities to Outcomes
- III. Developing Impact Evaluation Questions
- IV. The Counterfactual and Attribution
- V. Overview of Impact Evaluation Designs
- VI. Steps in Conducting an Impact Evaluation
- VII. Private-Sector Case Exercise
- VIII. Interpretation, Management Use, and Key Takeaways

Introduction to Impact Evaluation

Section I

Monitoring vs Evaluation

Monitoring ...

- is a continuous or regular collection and analysis of information about implementation to review progress;
- compares actual progress with what was planned so that adjustments can be made in implementation; and
- is an internal activity that is the responsibility of those who manage implementation. Monitoring thus represents a good management practice.

Other Characteristics of Monitoring:

- ✓ Program objectives are **not questioned**
- ✓ Indicators of progress are deemed to be appropriate
- ✓ Data collected routinely
- ✓ Does not address “what if” questions

*Note the **nonjudgmental** nature of monitoring. With monitoring, a judgment has already been made that the project or program should be implemented.*

Monitoring vs Evaluation

Evaluation ...

- is a periodic assessment of the relevance, efficiency, effectiveness, impact, and sustainability of a project, program, or intervention
- is the systematic, rigorous, and impartial assessment of a program or project to ascertain its level of achievement. (NEPF Guidelines)

Key Features of Evaluation:

- “measuring effects” requires a methodology for the evaluation
- “the effects” emphasize outcomes and impacts
- “comparison of effects with goals” requires explicit criteria for judging whether goals and objectives have been achieved.

Why evaluate?

- Fight the 3 Is
 - **Ideology:** Why do we run a customer loyalty program?
Because customers love rewards
 - **Ignorance:** How do we measure the impact of our CSR program? *Number of trees planted and pesos donated*
 - **Inertia:** Why do we use this sales incentive scheme?
Because it is the same scheme we have used for the last 10 years

What are the risks when we don't evaluate?

- Waste company resources
- Missed opportunities
- No proper attribution of positive changes
- No proper guide post on how to move forward

Why evaluate?

- We need to learn (**learning objective**)
 - What works and what doesn't
- We need to manage better (**management objective**)
 - How can we make our products, programs, and processes more efficient and effective?
- We need to answer to the board, investors, and customers (**accountability objective**)
 - How can we justify continuing, expanding, or cutting existing programs?

What is impact evaluation (IE)?

- The impact of a program is defined as a comparison between:
 - a. What actually happens after the program has been introduced
 - b. What would have happened had the program not been introduced (i.e., the “counterfactual”)
- These 2 ingredients are required to demonstrate impact

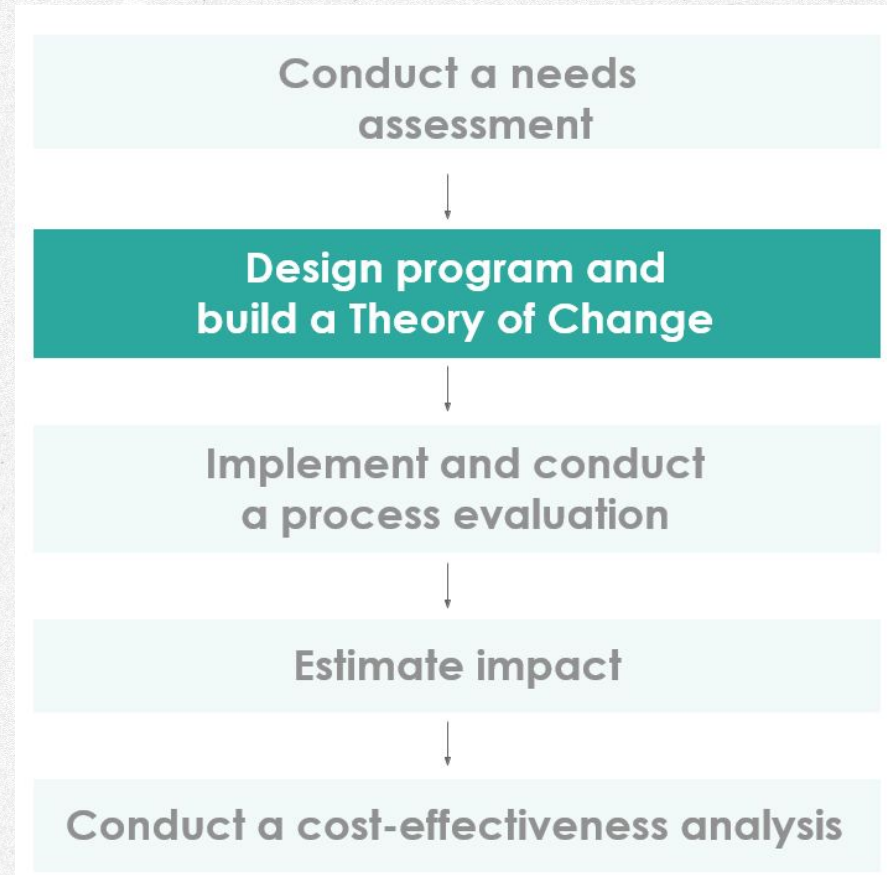
One program, four lenses

Type	Question it answers	Example: sales training for branch staff	What it cannot tell you
Routine monitoring	Are we doing what we planned?	No. of staff trained, sessions held, budget spent	<i>Whether the training changed anything</i>
Process evaluation	Is the program delivered well?	Trainer quality, attendance, participant feedback	<i>Whether sales rose because of the training</i>
Outcome tracking	Did the target indicator move?	Average monthly sales per staff before and after the training	<i>Whether sales would have risen anyway</i>
Impact evaluation	How much of the change did the program cause?	Sales of trained staff vs. a comparable group of untrained staff	<i>Only as credible as the comparison group</i>

Theory of Change: From Activities to Outcomes

Section II

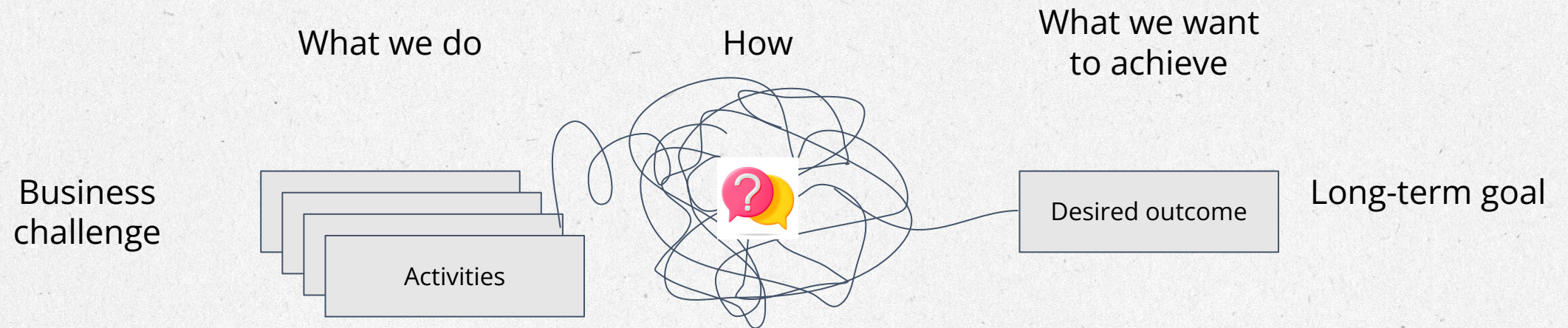
Foundation of impact evaluation - TOC



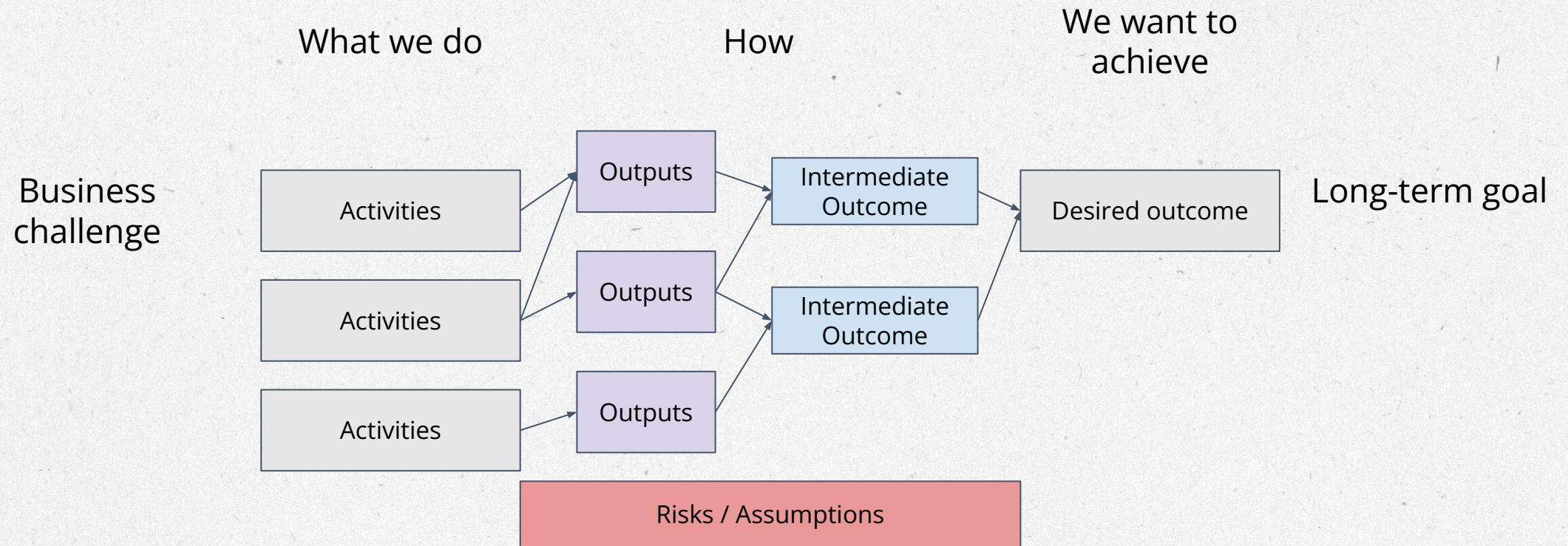
What is a *theory of change (TOC)*?

- Describes **how** and **why** a particular intervention will deliver the desired results or intended outcomes
- Details the **logical steps** to achieve the outcomes
- Outlines the activities that will generate outputs that lead to outcomes
- Shows the **causal relationship** and assumptions and risks

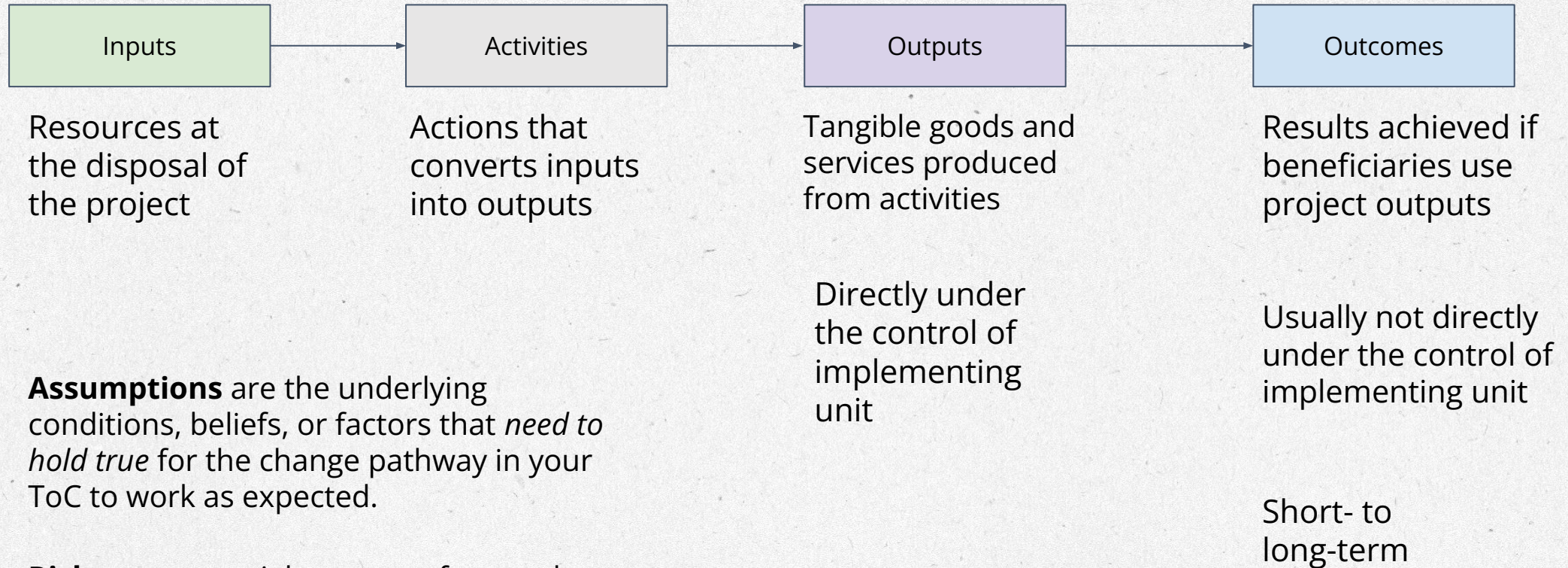
From business challenge to outcomes



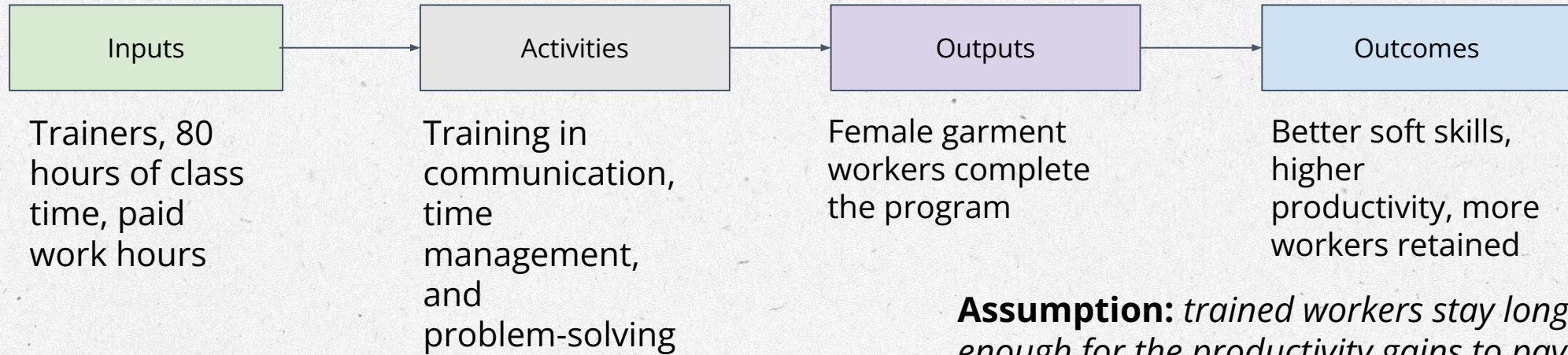
From business challenge to outcomes



Building a theory of change



Example: Soft-skills training at a garment exporter



Assumption: *trained workers stay long enough for the productivity gains to pay off*

Risk: *trained workers leave for better-paying jobs*

Result: In a randomized trial in five factories in Bengaluru, India, the net return was about 250% nine months after the program ended.

Outputs vs. outcomes: common mistakes

Program	Unit	Output Indicator	Outcome Indicator	Typical Mistake
Sales training for branch staff	HR / Sales	No. of branch staff who completed the training	Average monthly sales per staff, 6 months after training	<i>Reporting training completion as evidence that sales improved</i>
Customer loyalty app	Marketing	No. of app downloads; no. of active members	Repeat purchase rate of members within 12 months	<i>Treating sign-ups as proof that the app drives repeat buying</i>
CSR scholarship program	CSR / Foundation	No. of scholarships awarded; amount disbursed	% of scholars who graduate on time and find work within 1 year	<i>Reporting pesos disbursed as the program's impact</i>
Employee wellness program	HR	No. of employees who joined wellness activities	Sick days and health claims per employee	<i>Treating participation rates as evidence of healthier, more productive staff</i>

Developing Impact Evaluation Questions

Section III

TOC is important for...

- Designing the impact evaluation and interpreting the results
 - Generate research questions
 - Decide which outcomes to measure and what data to collect
 - If no effect: At what step did the hypothesized causal chain break?
 - If positive effect: What (do we believe) is the mechanism through which the program works?
 - forces us to think critically about the “why,” thus giving richer lessons for decision-makers and more generalizable knowledge

TOC and Evaluation Questions

- A well-formulated evaluation question keeps the IE study **focused and relevant**.
- It ensures the evaluation addresses the **key policy decisions** facing stakeholders.
- Broad or vague questions lead to unfocused data collection and analysis.
- Specific questions guide the choice of evaluation design, indicators, and data sources.

Characteristics of Good Evaluation Questions

- **Clear** and unambiguous wording that all stakeholders understand.
- **Directly linked** to the program's theory of change and intended outcomes.
- **Feasible to answer** given available time, resources, and data.
- **Action-oriented** so the answers can inform program or business decisions.

Examples of Impact Evaluation Questions

- “What was the effect of the sales training on monthly sales per staff?”
- “Did the loyalty app change how often customers come back?”
- “Which customer segments responded most to the promotion?”

Discussion: which of these are good evaluation questions?

1. How many employees attended the leadership training?
2. Which customer segments increased their spending the most after the price change?
3. Why are sales so low in our Visayas branches?
4. What is management's plan to improve employee retention?
5. Did the new onboarding program reduce first-year resignations?

TOC and Indicators

- Indicators **measure progress** toward outputs, outcomes, and impacts identified in the theory of change.
- They provide a way to track whether the program is on course to achieve its objectives.
- Performance indicators focus on activities and outputs; outcome indicators capture changes attributable to the intervention.
- Both are needed for a complete picture of program performance.

Validity and reliability: choosing good indicators

Validity: measuring the right thing

- Does the indicator truly capture the outcome in our TOC?
- Can we collect the data without bias?

Example: *loyalty app downloads say little about whether customers keep coming back.*

Reliability: measuring it consistently

- Would we get the same result if we measured again, or if someone else did?
- How time- or situation-specific is the indicator?

Example: *staff ratings without a rubric change with whoever does the rating.*

When choosing indicators

Match the specific change in your TOC, use clear definitions and tested instruments, and weigh cost against quality without giving up credibility.

Validity and reliability: examples

Where it lands	Public program example	Business example	How to fix it
Reliable, not valid	Brochures distributed, as a measure of health awareness	Training sessions held, as a measure of sales skills	<i>Measure the change itself: knowledge, skills used, sales</i>
Valid, not reliable	Teacher quality rated by principals with no standard criteria	Staff performance rated by supervisors with no rubric	<i>Clear definitions, scoring rubrics, trained raters</i>
Neither	Water quality judged by households from taste and color	Customer satisfaction judged from front-line staff impressions	<i>Lab tests; standard customer surveys</i>
Valid and reliable	Child height-for-age from standard measurements	Monthly sales per staff from the POS system	

The Counterfactual and Attribution

Section IV

Correlation is not causation

- Causally unrelated variables might happen to be highly correlated with each other
- Causally related variables might have non-linear relationships

Examples:

- Crime rate is higher during months of peak ice cream sales
- Children with bigger shoe sizes tend to read better
- Coffee consumption and alertness
- Exercise and physical performance

Reverse causality

- Even if there is a causal relationship, the direction can be unclear

Examples:

- presence of traffic enforcer and severity of traffic
- perseverance/diligence and poverty
- ad spending and sales (firms spend more on ads when they expect high sales)



**THEY ARE POOR
BECAUSE THEY ARE LAZY**

TRUE

FALSE

The basic impact evaluation formula

The basic impact evaluation question:

What is the impact or causal effect of a program (P) on an outcome of interest (Y)?

$$\Delta = (Y \mid P = 1) - (Y \mid P = 0)$$

Δ : Causal impact of the program

Y : Outcome of interest

P = 1 : With the program (treatment)

P = 0 : Without the program (comparison)

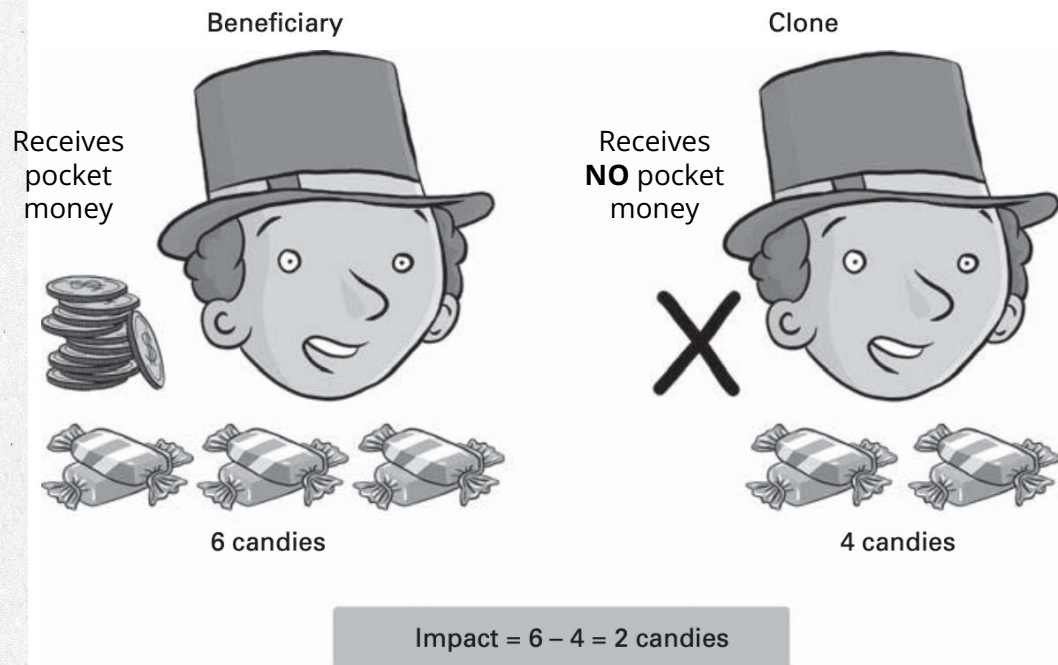
The Counterfactual Problem

The **counterfactual** is what would have happened, that is, what the outcome (Y) would have been for a program participant in the absence of the program (P).

In the basic impact evaluation formula, $(Y \mid P = 0)$ represents the counterfactual.

Hypothetical Solution: If we could find a *perfect clone* for each participant

Figure 3.1 The Perfect Clone



But, unfortunately, perfect clones don't exist

- Even identical twins have important differences
- Exact replication is impossible

Since we cannot directly observe the counterfactual, we must estimate it.

Estimating the counterfactual: From individual to group

- Although no perfect clone exists for a single unit, we can use statistical properties to create **statistically identical groups**
- Two groups
 - Treatment group: Participate in the program, outcome:
 $(Y \mid P = 1)$
 - Comparison group: Remains unaffected by program, estimates
 $(Y \mid P = 0)$
- Identify treatment and comparison groups that are statistically identical, on average, in the absence of the program.

“Counterfeit” estimates of the counterfactual

- Before-and-after comparison
 - Also known as pre-post or reflexive comparison
 - Compare outcomes of same group before and after program
 - Assumes counterfactual = baseline outcome
- Enrolled-and-nonenrolled comparisons
 - Also known as self-selected comparisons
 - Compare participants with non-participants
 - Ignores selection bias

These are common but risky methods that often lead to inappropriate estimates of the counterfactual and can lead to severely biased impact estimates.

Why are these problematic?

Problems with before-and-after comparisons

The flawed assumption

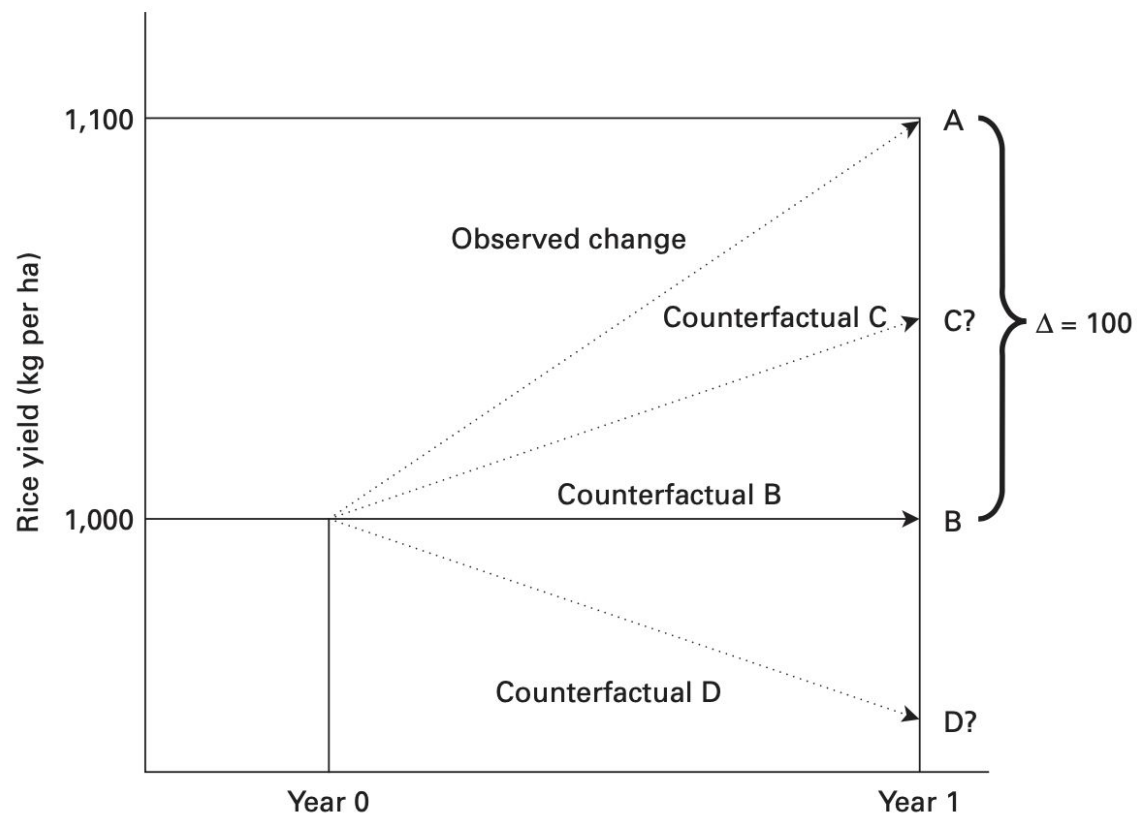
Before-and-after comparisons assume that without the program, outcomes would remain exactly the same as before the program started.

Example: Microfinance program that provides microloans for farmers to buy fertilizer

- Year 0 (baseline): Yield of 1,000 kg rice per hectare
- Year 1 (after the program): Yield of 1,100 kg rice per hectare
- Before-and-after estimate: 100 kg increase

What could go wrong?

Figure 3.3 Before-and-After Estimates of a Microfinance Program



Note: Δ = Change in rice yield (kg); ha = hectares; kg = kilograms.

Before-and-after estimate: A-B

Case 1: Normal rainfall in the year before the program was launched, but a drought occurred during the program

- What is the counterfactual?
- What would have been the estimate?

Case 2: Rainfall improved during the program

- What is the counterfactual?
- What would have been the estimate?

Many factors affect outcomes over time.
The baseline is almost never a good estimate of the counterfactual.

Intervention	Outcome Measured	Bias Type	Why the Before-After Estimate May Be Wrong
Digital marketing campaign	Online orders 3 months later	History / Seasonal variation	<i>If the campaign ran over payday weeks or the holiday season, order growth may reflect the season, not the campaign.</i>
New store layout in all branches	Monthly sales per branch	History (concurrent events)	<i>Sales were already rising because of a price cut and a competitor's store closure. The before-after gap overstates the layout's contribution.</i>
Six-month enterprise development training program	Average monthly business revenue at endline	Attrition (survivor bias)	<i>Participants who dropped out, often the worst performers, are excluded from endline measurement. The remaining group looks more successful than the full cohort actually was.</i>

Problems with comparing enrolled and nonenrolled (self-selected) groups

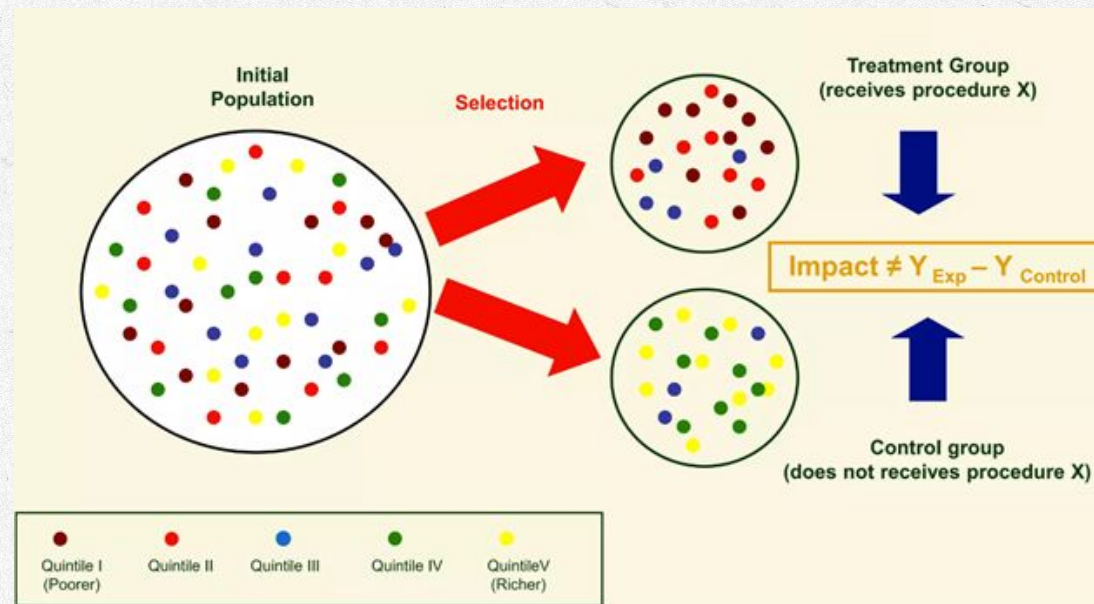
- Compares a group of individuals that voluntarily signs up for the program to a group that chooses not to participate
- A comparison group that self-selects out of a program provides a “counterfeit” estimate of the counterfactual

Example: vocational training program for unemployed youth

- Estimate impact on income by comparing the average incomes of a group of youth who chose to enroll to a group of youth who chose not to enroll despite being eligible to enroll
- What is the counterfactual? Is it a good estimate of the counterfactual?

Problems with comparing enrolled and nonenrolled (self-selected) groups

Selection bias occurs when the reasons for which an individual participates in a program are correlated with outcomes. Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation.



Intervention	Outcome Measured	Bias Type	Why the With-Without Comparison May Be Wrong
Microfinance loan program for small businesses	Business revenue of borrowers vs. non-borrowers	Self-selection	<i>Entrepreneurs motivated and confident enough to apply for loans are already more capable. Higher revenue may reflect who they are, not what the loan did.</i>
CSR scholarship program for college students	Graduation rates of recipients vs. non-recipients	Cream-skimming (selection on merit)	<i>Scholarships typically go to high-achieving students who would likely have graduated anyway.</i>
Customer loyalty program	Spending of members vs. non-members	Self-selection	<i>Customers who join are already frequent, high-spending shoppers. Their higher spending may reflect who they are, not what the program did.</i>

A counterfeit counterfactual at work: employee wellness

Illinois Workplace Wellness Study

- About 5,000 university employees were randomly offered a wellness program or not
- Comparing participants with non-participants: participants went to the gym 7.4 times a year, non-participants 3.8 times
- The comparison found little to no effect on health behaviors or medical spending in the first year

Lesson

Employees who chose to join were already healthier and spent less on health care before the program. The participant vs. non-participant gap was mostly selection, not impact.

A counterfeit counterfactual at work: paid search ads

eBay paid search experiments

- Customers who click on search ads buy more, so the ads looked highly profitable
- eBay switched off ads on searches for its own brand (e.g., “ebay shoes”) in a controlled test
- Nearly all of that traffic came in through the free search results instead
- Returns from paid search were a fraction of the non-experimental estimates

Lesson

Clicking an ad and buying are both driven by intent to buy. Without a counterfactual, the ads got credit for sales that would have happened anyway.

Overview of Impact Evaluation Designs

Section V

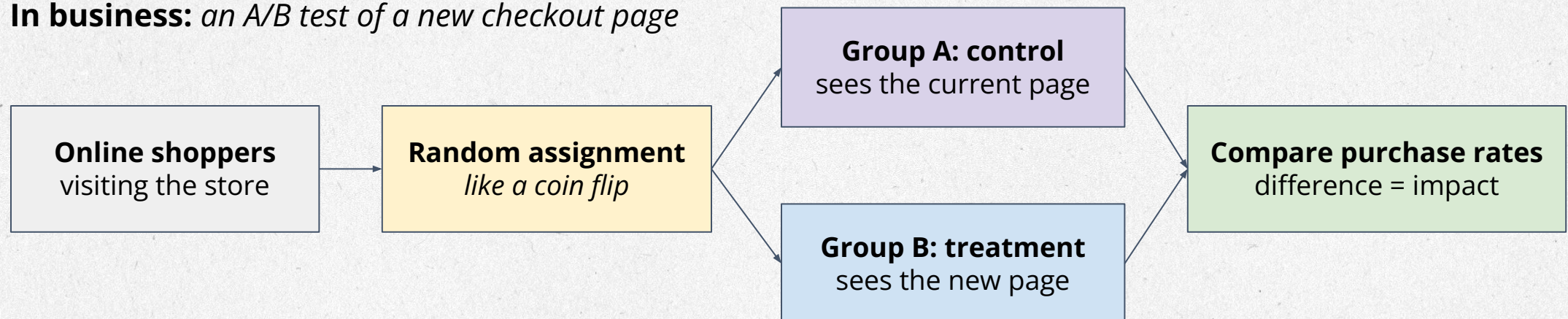
Randomized Control Trial (RCT)

- An impact evaluation method in which all **eligible units** in a **sample** are **randomly assigned** to **treatment** and **control** groups.
- In business, an **A/B test** is an RCT: customers, staff, or branches are randomly assigned to version A or version B.

Solution to the counterfactual problem

- **Random assignment** of treatment ensures that the treatment group and the control group **are similar in all observable and unobservable characteristics before the intervention.**
- Control group = statistical twin of the treatment group
- Any systematic differences in the outcomes of interest between the two groups are solely due to the treatment.

In business: *an A/B test of a new checkout page*



RCTs in the private sector

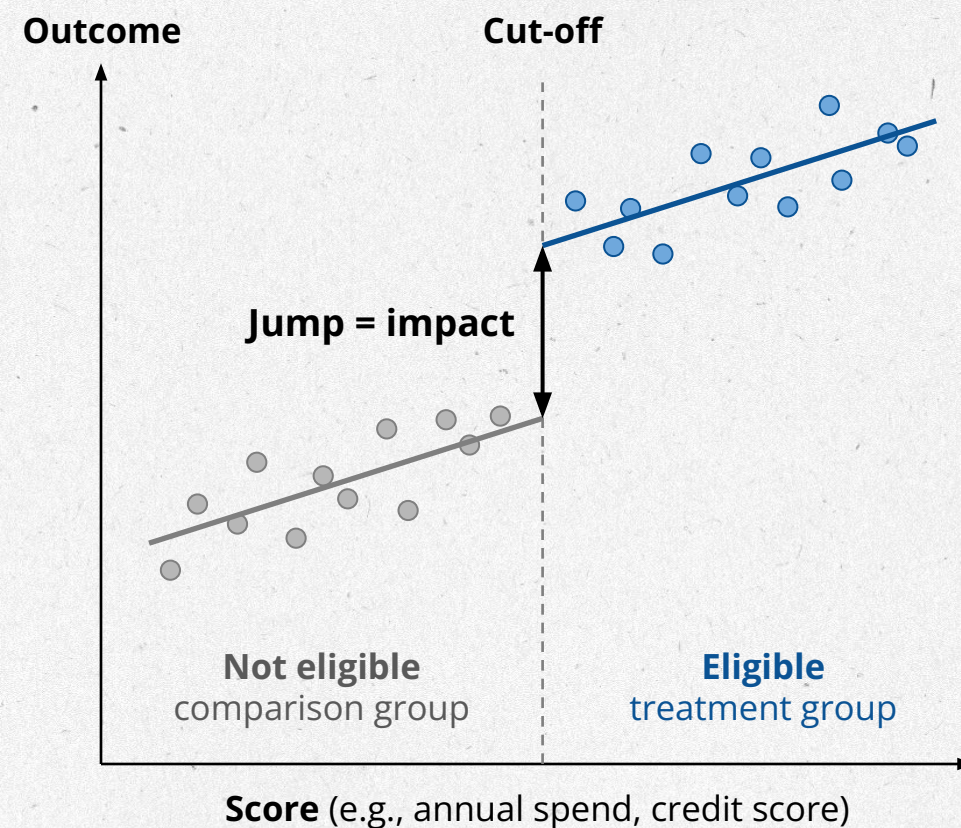
Business function	Examples of company RCTs
Marketing & sales	A/B tests of websites, prices, promos, and ad formats (Microsoft Bing, eBay)
Human resources	Work-from-home trial for call-center staff at Ctrip, China: performance up 13%, attrition halved
Operations	Free management consulting for randomly chosen textile plants in India: productivity up 17% in the first year
Financial services	First Macro Bank, Manila: approval randomized for marginal loan applicants to measure the effect of credit
Training & CSR	Soft-skills training at Shahi Exports garment factories, India: about 250% net return

Quasi-experimental designs: when we cannot randomize

- The missing ingredient is random assignment; the key issue is finding a valid counterfactual
- Problems in non-randomized settings:
 - Non-random program participation (self-selection)
 - Non-random program placement (programs are targeted)
 - Non-random attrition or drop-out
- Typical uses: when we cannot randomize, or when evaluating past or ongoing programs
- Critical questions: What is the counterfactual group? Why is it credible?

Regression Discontinuity Design (RDD)

- Used when eligibility is set by a cut-off on a continuous score (e.g., credit score, annual spend, test score)
- Units just above and just below the cut-off are similar in all respects except treatment
- **The outcomes of those just below the cut-off serve as the counterfactual for those just above it**
- The jump in outcomes at the cut-off is the program impact
- Main risk: people who can adjust their score to land on the preferred side



RDD in business

Example	Score (running variable)	Cut-off	What we learn
Loan approval at a bank	Applicant credit score	Minimum score for approval	<i>Effect of getting a loan on business growth, for applicants near the threshold</i>
Customer loyalty tiers	Annual spend	Spend needed for “Gold” status	<i>Effect of Gold perks on spending, for customers near the threshold</i>
Online restaurant ratings (Yelp, San Francisco)	Average review score	Rounding up to the next half-star (e.g., 3.74 vs. 3.76)	<i>Moving from 3 to 3.5 stars raised the chance of selling out at peak hours from 13% to 34%</i>

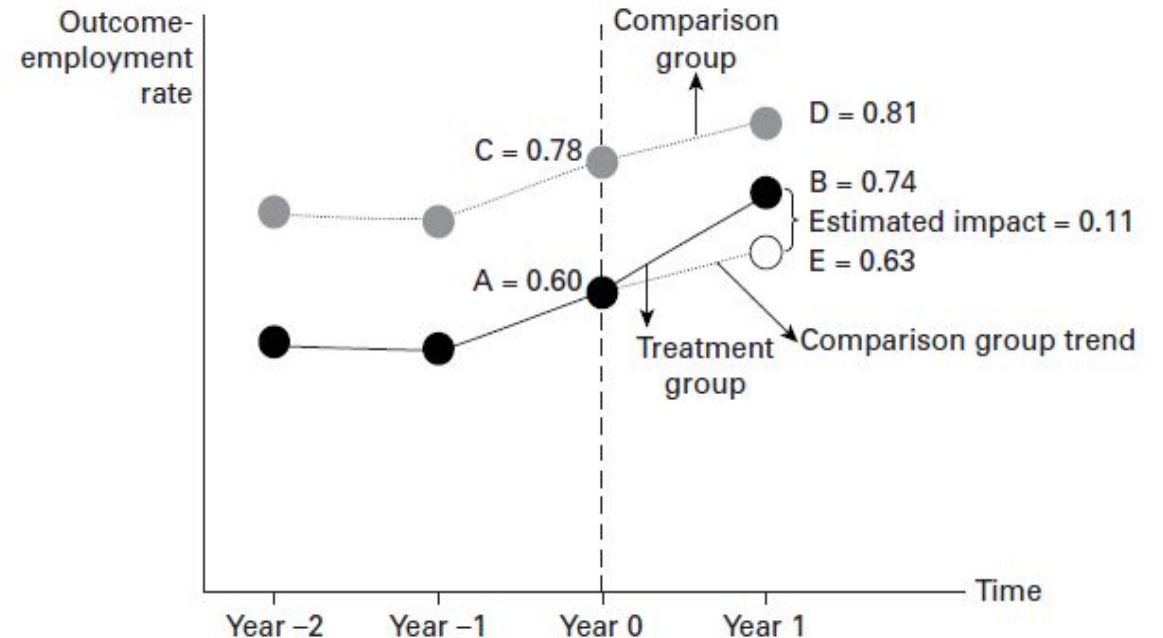
Difference in Difference (DID)

- a quasi-experimental design that estimates impact by comparing the **change in outcomes over time** for a group exposed to the intervention (treatment) against the change for a group that is not (comparison).
- Example: a new pricing policy starts in Luzon branches first. Compare the change in Luzon sales with the change in Visayas and Mindanao sales over the same period.

DID requires parallel trends

- Counterfactual argument: The missing counterfactual is provided by a comparison group that is **evolving similarly** prior to and after the treatment;

Figure 7.1 The Difference-in-Differences Method



DID requires parallel trends

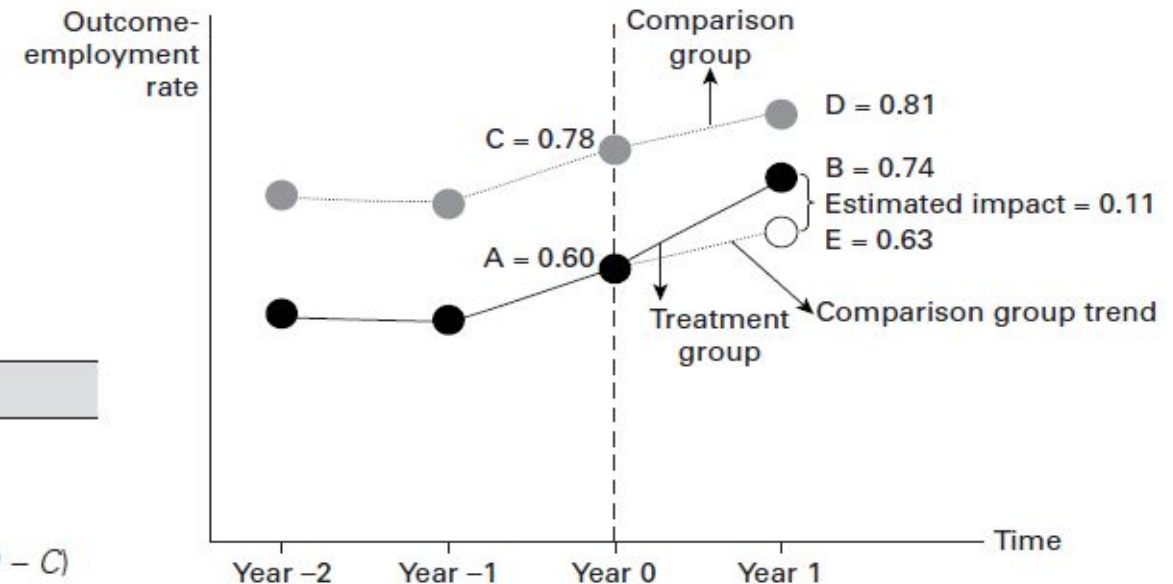
- Treatment effect: (Difference in outcome of treated group before and after the program) - (Difference in outcome of control group before and after the program)

Table 7.1 Calculating the Difference-in-Differences (DD) Method

	After	Before	Difference
Treatment/enrolled	B	A	$B - A$
Comparison/nonenrolled	D	C	$D - C$
Difference	$B - D$	$A - C$	$DD = (B - A) - (D - C)$

	After	Before	Difference
Treatment/enrolled	0.74	0.60	0.14
Comparison/nonenrolled	0.81	0.78	0.03
Difference	-0.07	-0.18	$DD = 0.14 - 0.03 = 0.11$

Figure 7.1 The Difference-in-Differences Method



Propensity Score Matching (PSM)

- A quasi-experimental method that:
 - Estimates each unit's propensity score, the probability of receiving the program given a set of observed baseline characteristics; and
 - Pairs (or weights) treated units with comparison units that have similar scores, creating a control group that is, on observables, statistically comparable to the treated group.
- **Treatment effect: difference in outcomes between matched treatment and control units**
- Assumes every factor that influences both program take-up and outcomes is observed and included in the score.

Propensity Score Matching (PSM)

When to use PSM?

- **Participation is voluntary or admin-selected on *observable* traits.**
If take-up hinges on things you can measure (wealth, distance, education) and not on hidden motivation, PSM is a fit.
- **No clean cut-off, no staggered roll-out, but good overlap.**
Programs allocated by case-worker discretion or open enrolment lack an RDD cut-off and may start everywhere at once, leaving cross-sectional matching as the main quasi-experimental option.

Steps in Conducting an Impact Evaluation

Section VI

Prospective vs Retrospective Impact Evaluation

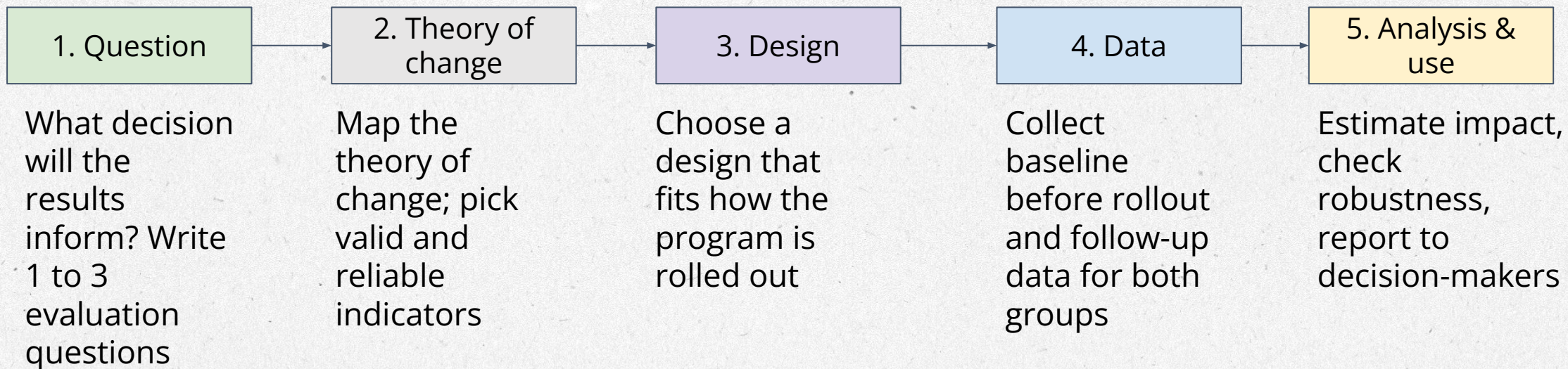
Prospective evaluation

- Developed at the same time as the program is being designed and built into program implementation
- Baseline data are collected before the program is implemented for both the group receiving the intervention and the group used for comparison that is not receiving the intervention

Retrospective evaluation

- Assess program impact after the program has been implemented, looking for treatment and comparison groups ex post

The impact evaluation roadmap



Start early: *plan the evaluation while the program is being designed. This keeps the most design options open.*

Step 1: Define the question



- **Start from the decision:** what will management do with the answer? Scale up, redesign, or stop?
- **Write 1 to 3 causal questions:** what is the effect of [program] on [outcome] for [whom], compared with no program?
- **Define the population and unit:** customers, staff, branches, or suppliers, and who is eligible
- **Agree on what success looks like:** the main outcome and the smallest effect worth acting on
- **Secure buy-in and budget early:** the people running the program must agree on how it will be rolled out

Step 2: Map the theory of change



- **Lay out the pathway:** activities, outputs, and outcomes, with the assumptions that link them
- **Pick indicators for each level:** primary and secondary outcomes that are valid and reliable
- **Track the steps in between:** take-up and usage show where the chain breaks if the impact turns out small
- **Set the timing:** when should each outcome respond? Schedule follow-up to match the theory of change
- **Anticipate side effects:** spillovers to the comparison group, or customers shifting purchases instead of buying more

Step 3: Choose the design



- **Let the rollout rule guide the method:**
 - **RCT:** a lottery, pilot, or phased rollout is possible
 - **RDD:** eligibility depends on a cut-off score
 - **DID:** the program starts at different times in different branches or areas, and pre-program data exist
 - **PSM:** sign-up is voluntary, and we have rich data on joiners and non-joiners
- **Choose the unit of assignment:** individuals give more statistical power; branches or areas limit spillovers but need more units
- **Define what the comparison group gets:** the current product, nothing new, or the program later
- **Check the sample size:** enough units to detect the smallest effect worth acting on, plus a buffer for drop-outs

Step 4: Collect the data



- **Collect baseline data before rollout:** check that the groups look alike on key characteristics (balance)
- **Measure both groups the same way:** same instruments, definitions, and timing in every round
- **Use administrative data where possible:** sales, account, and HR records are cheaper and often more reliable than surveys
- **Monitor implementation:** record who actually received the program and whether the comparison group stayed untreated
- **Hold on to the sample:** track units to limit drop-outs, and follow the Data Privacy Act on consent and data handling

Step 5: Analyze and use the results



- **Estimate the impact:** the difference in outcomes between the treatment and comparison groups, with a confidence interval
- **Test robustness:** re-run with another comparison group, or with an outcome the program should not affect
- **Report the correct impact parameter:** Average Treatment Effect (ATE), Intent to Treat Effect (ITT), Average Treatment Effect on the Treated (ATT), among others
- **Look at subgroups only if planned:** for example, new vs. existing customers
- **Link the result to the decision:** compare impact with cost, then scale, redesign, or stop

Guidelines in timing evaluations

- **Not too early:** you need to allow enough exposure for the program's expected pathways to operate
- **Not too late:** do not wait so long that other policies, market shocks, or high attrition creep in and you lose statistical power, and you risk missing the decision window in which managers must decide on scale-up or redesign.

How to Budget for an Impact Evaluation

- **Know the big cost drivers first**
 - Field data collection (enumerator wages, training, transport, tablets) typically **60–70 percent of the entire evaluation budget**
 - Sample size, number of survey rounds, and travel distance multiply those costs; trimming just 10 percent of the sample often frees more money than cutting any other line item.

In-house or outsourced evaluation?

Ownership vs. Independence Trade-off

- Choose a model that balances management ownership (to ensure uptake) with analytic independence (to protect credibility).

Capacity & Timeline Assessment

- Match the model to internal statistical skills, procurement rules, and the decision calendar. Fast decisions often favor in-house or embedded teams.

Case Walkthrough: Green Bank of Caraga

Section VII

The case: Green Bank of Caraga

Context

- A rural bank in Butuan City wants its clients to save more
- Many clients want to save but find it hard to keep money in the account
- The bank designs SEED (Save, Earn, Enjoy Deposits): clients set a goal amount or date and cannot withdraw until they reach it
- Same interest rate as a regular savings account; bank staff visit clients to offer it

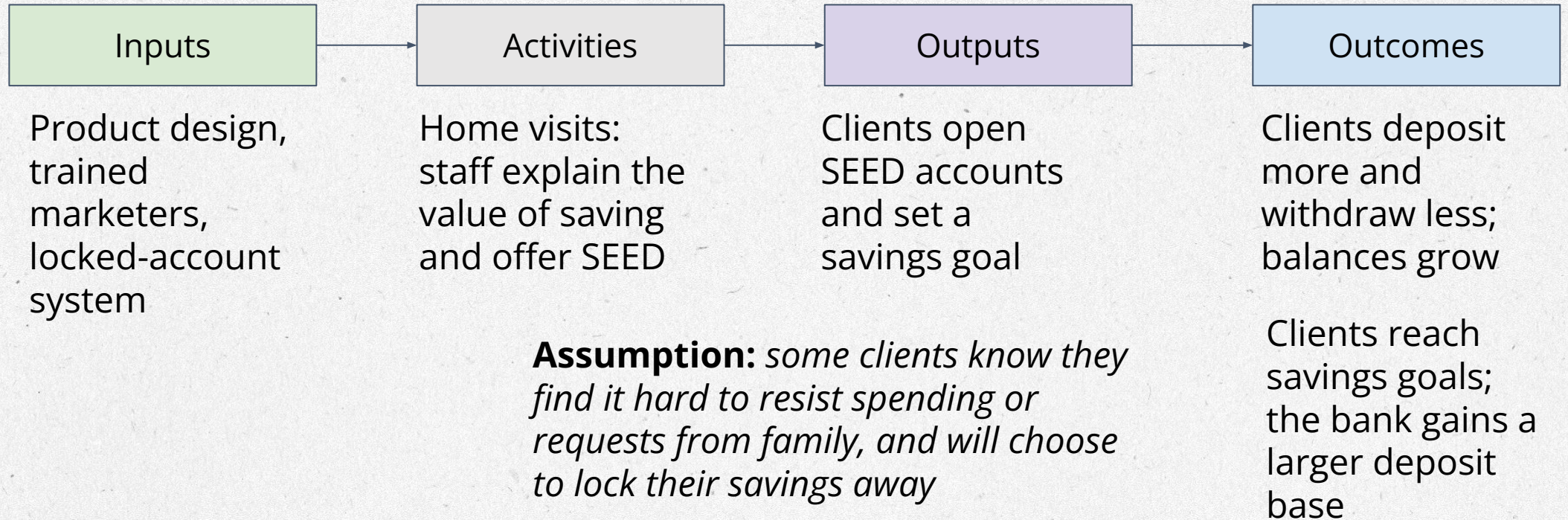
The bank's question

Before offering SEED in all branches, management wants to know: does SEED actually make clients save more?

Guide questions for discussion

1. What is SEED's theory of change, from inputs to outcome?
2. What impact evaluation question should the bank ask, and which indicators would answer it?
3. Why would comparing SEED users with non-users, or clients' savings before and after, mislead?
4. Which evaluation design would give the bank a credible answer?

Q1. SEED's theory of change



Risk: *clients only move money from other accounts, so total savings do not rise*

Guide questions for discussion

1. What is SEED's theory of change, from inputs to outcome?
2. What impact evaluation question should the bank ask, and which indicators would answer it?
3. Why would comparing SEED users with non-users, or clients' savings before and after, mislead?
4. Which evaluation design would give the bank a credible answer?

Q2. Evaluation question and indicators

Evaluation question

Did offering SEED increase clients' savings, compared with what they would have saved without it?

Outcome indicators

- Take-up: share of clients offered SEED who opened an account
- Average savings balance at the bank, 6 and 12 months after the offer
- Total savings, including savings held outside the bank

Guide questions for discussion

1. What is SEED's theory of change, from inputs to outcome?
2. What impact evaluation question should the bank ask, and which indicators would answer it?
3. Why would comparing SEED users with non-users, or clients' savings before and after, mislead?
4. Which evaluation design would give the bank a credible answer?

Q3. Why simple comparisons would mislead

Comparison	What it would show	Why it misleads
SEED users vs. non-users	Users have higher balances	
Clients' savings before vs. after SEED	Balances grew over the year	
Clients picked by staff for SEED vs. all other clients	Higher savings among those picked	

Q3. Why simple comparisons would mislead

Comparison	What it would show	Why it misleads
SEED users vs. non-users	Users have higher balances	<i>Clients who choose a locked account are likely keen savers already (self-selection). The gap reflects who they are, not what SEED did.</i>
Clients' savings before vs. after SEED	Balances grew over the year	
Clients picked by staff for SEED vs. all other clients	Higher savings among those picked	

Q3. Why simple comparisons would mislead

Comparison	What it would show	Why it misleads
SEED users vs. non-users	Users have higher balances	<i>Clients who choose a locked account are likely keen savers already (self-selection). The gap reflects who they are, not what SEED did.</i>
Clients' savings before vs. after SEED	Balances grew over the year	<i>Balances can grow anyway: harvest income, remittances, a raise, or other bank promotions (history).</i>
Clients picked by staff for SEED vs. all other clients	Higher savings among those picked	

Q3. Why simple comparisons would mislead

Comparison	What it would show	Why it misleads
SEED users vs. non-users	Users have higher balances	<i>Clients who choose a locked account are likely keen savers already (self-selection). The gap reflects who they are, not what SEED did.</i>
Clients' savings before vs. after SEED	Balances grew over the year	<i>Balances can grow anyway: harvest income, remittances, a raise, or other bank promotions (history).</i>
Clients picked by staff for SEED vs. all other clients	Higher savings among those picked	<i>Staff may pick their best customers. The home visit itself may also raise savings, apart from the product.</i>

Guide questions for discussion

1. What is SEED's theory of change, from inputs to outcome?
2. What impact evaluation question should the bank ask, and which indicators would answer it?
3. Why would comparing SEED users with non-users, or clients' savings before and after, mislead?
4. Which evaluation design would give the bank a credible answer?

Q4. What the bank did: an RCT with three groups

Group	What clients received	Share of the 1,777 clients	What the comparison tells us
Commitment group	Home visit about saving, plus an offer to open a SEED account	One half	<i>Effect of offering SEED (vs. control)</i>
Marketing group	The same home visit and savings message, but no SEED offer	One quarter	<i>Effect of the visit alone</i>
Control group	Nothing new	One quarter	<i>The counterfactual</i>

What the bank found after one year

Results

- 28% of clients offered SEED opened an account
- Average savings balances in the commitment group rose 42% after six months and 82% after one year, compared with the control group
- The increase did not come at the expense of savings held elsewhere

Lessons for managers

The effect is measured on everyone offered SEED, including the 72% who said no. The marketing group shows how much came from the product rather than the visit.

Interpretation, Management Use, and Key Takeaways

Section VIII

IE designs side by side

Criterion	RCT	RDD	DiD	PSM
Use when	We can assign the program by lottery, pilot, or phased rollout	Eligibility depends on a cut-off score	Rollout is staggered across places or time, and pre-program data exist	Sign-up is voluntary and we have rich data on who joins
Comparison group	Units randomly assigned not to get the program (yet)	Units just below the cut-off	Untreated units, over the same period	Non-participants matched on observed characteristics
Key assumption	Groups stay as assigned; no spillovers	No one can manipulate the score; units near the cut-off are alike	Without the program, both groups would follow parallel trends	Everything that drives both joining and outcomes is observed
Data needed	Baseline and follow-up; assignment list	Score, cut-off, and outcomes on both sides	At least three time points for both groups	Rich baseline data on participants and non-participants

From evidence to decision

Company	Program evaluated	What the evaluation found	Implication for management
Ctrip (China)	Work from home for call-center staff	Performance up 13%; attrition halved	<i>Scale: offered work from home company-wide; gains rose to 22% once staff could choose</i>
eBay (US)	Ads on searches for its own brand name	No measurable short-term benefit	<i>Rethink: focus ad spending on new and infrequent customers, who did respond to ads</i>
University employer (US)	Workplace wellness program	Little to no effect on health or medical spending in year 1	<i>Redesign or temper expectations; keep measuring before expanding</i>
Opower and US utilities	Home energy reports	Effects at the first 10 sites overstated effects at the next 101	<i>Scale with caution: early pilot sites may not represent later ones</i>

Key takeaways

- Monitoring tells us whether we did what we planned; impact evaluation tells us whether it made a difference
- A theory of change links activities to results and points to the right questions and indicators
- **The impact estimate is only as good as the counterfactual;** before-after and participant vs. non-participant comparisons often mislead
- Choose the design that fits how the program is rolled out, and plan it before launch
- Use the results to decide: continue, scale, redesign, or stop

References

- Adhvaryu, A., Kala, N., and Nyshadham, A. 2023. Returns to On-the-Job Soft Skills Training. *Journal of Political Economy* 131(8).
- Allcott, H. 2015. Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics* 130(3).
- Anderson, M., and Magruder, J. 2012. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *Economic Journal* 122(563).
- Ashraf, N., Karlan, D., and Yin, W. 2006. Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines. *Quarterly Journal of Economics* 121(2).
- Blake, T., Nosko, C., and Tadelis, S. 2015. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica* 83(1).
- Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. 2013. Does Management Matter? Evidence from India. *Quarterly Journal of Economics* 128(1).
- Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. 2015. Does Working from Home Work? Evidence from a Chinese Experiment. *Quarterly Journal of Economics* 130(1).
- Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., and Vermeersch, C. M. J. 2016. *Impact Evaluation in Practice*, Second Edition. World Bank.
- Jones, D., Molitor, D., and Reif, J. 2019. What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study. *Quarterly Journal of Economics* 134(4).
- J-PAL. 2024. Lectures: Why Evaluate; Theory of Change and Measurement. Abdul Latif Jameel Poverty Action Lab, Cambridge, MA.
- Karlan, D., and Zinman, J. 2011. Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. *Science* 332(6035).
- Kohavi, R., and Thomke, S. 2017. The Surprising Power of Online Experiments. *Harvard Business Review*, September-October 2017.
- NEDA. 2020. Guidelines on Evaluation in the National Government.

Thank you!

Philippine Statistical Association, Inc.

2/F Philippine Social Science Center, Commonwealth Avenue, Diliman, Quezon City

(632) 8920-6513 | psai.secretariat@gmail.com | www.psai.ph