Unit II: Models and Designs of Evaluation

Presented by:
Pratiksha Thapa
Roll.no: 19
Aruna Subedi
Roll no: 8
Ranjita kumari Bohara
Roll no: 23
Anita Tiwari
Roll no: 5
Aakriti Mahat
Roll no: 1

Presented to: Associate Prof. Dr Dipendra Kumar Yadav, SHAS

Evaluation models

Evaluation models are systematic frameworks or approaches that guide how a program, project, or policy should be evaluated

They provide structured ways to;

- identify what to evaluate
- collect information
- judge performance
- interpret results

Evaluation models helps us organize and conduct an evaluation in a planned, logical, and meaningful way.

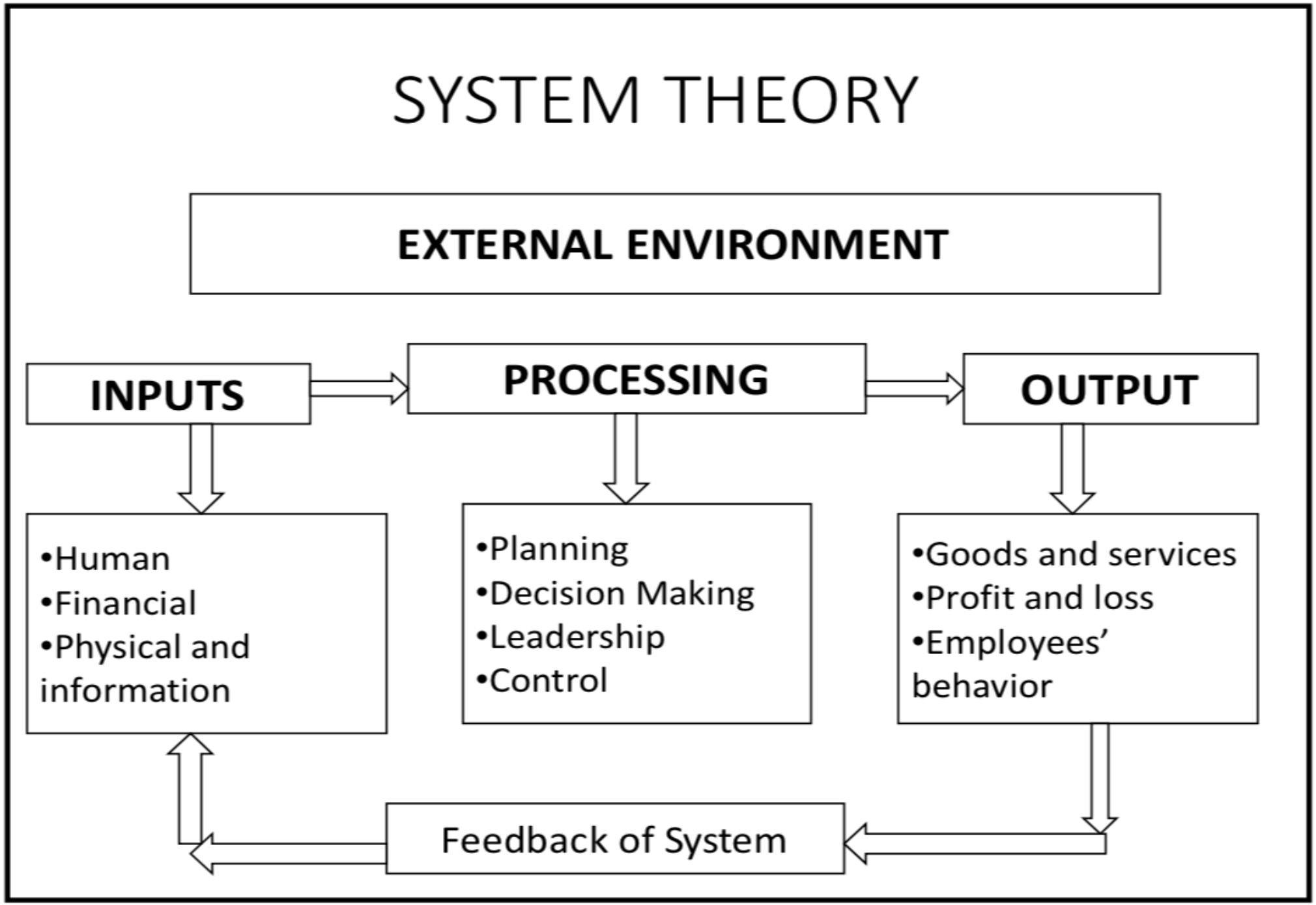
Four types of evaluation models

1. System analysis model
2. Behavioral objectives/ Goal- attainment/ Goal-based model
3. Decision making model
4. Goal- free model

System Analysis Model

- The systems analysis model developed by Rivlin (1971)
- System analysis is a problem solving technique that breaks down a system into its component/ pieces for the purpose of the studying how well those component parts work and interact to accomplish their purpose.
- Evaluates a program as a system made of inputs, processes, outputs, and feedback
- This model looks at the efficiency and effectiveness of the program

SYSTEM THEORY



Key Components

1. Input:

- elements that come into the system
- for example, resources used: manpower, funds, materials, technology.

2. Process:

- consists of things that occur as the program operates
- for example, activities performed: training, service delivery, communication, monitoring

3. Output:

- the results of the program
- for example, immediate results of activities: number vaccinated, number trained, number of IEC sessions held

4. Outcome/ Impact:

- long term changes in health status
- for example, Reduced mortality, improved behavior, increased service utilization

5. Feedback:

- information about performance used to improve the system.
- helps in continuous monitoring.

The evaluator examines the program's efficiency in light of these categories.

This model might be employed to determine whether a program is getting people into and out of the program in an efficient manner, as well as whether the program is achieving its goals.

Strengths

- Simple and logical
- Shows complete pathway of program functioning
- Useful for monitoring and evaluation planning and indicator development
- Widely applicable in public health programs

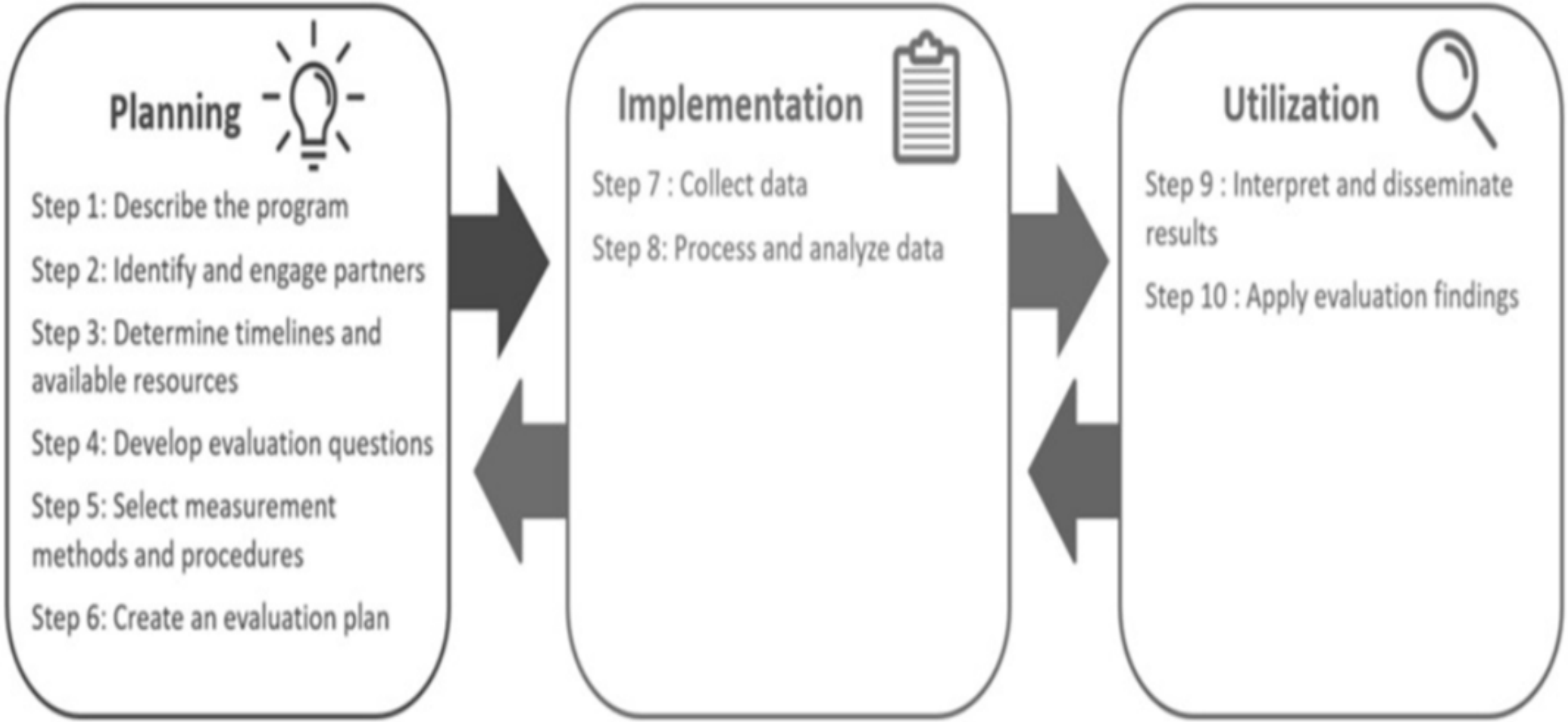
Limitations

- Does not deeply analyze why the outcomes occur
- Assumes linear cause effect relationship
- Less attention to external factors

Behavioral objectives/ Goal-attainment/ Goal-based Model

- This model was developed by : Ralph tyler (Tylerian model)
- Focus:
  - Clearly defined behavioural objectives
  - Measuring the extent to which objectives are attained
  - Often used in education, training, and health promotion programs
- Evaluates whether program goals and objectives are achieved
- The goals-based evaluation model consists of ten steps in three phases: planning, implementation and utilization.

The Ten Steps for Goals-Based Evaluation



Step 1: Describe the Program

- This step creates the foundation for evaluation.
- Clearly describe what the program is about—its purpose, goals, target population, activities, indicators, inputs, outputs, and expected results.
- Example,
  A school-based nutrition program aims to reduce junk-food consumption among grade 6–8 students by promoting healthy snacks.

Step 2: Identify and Engage Partners

- This step involves identifying all stakeholders (implementers, community leaders, teachers, local government, FCHVs, etc.) who will be impacted by the evaluation's implementation or results, and and involve them in evaluation planning. Consider both internal and external partners, including the program audience.
- Example,
  Engaging teachers, school principals, parents' groups, and local health office in the nutrition education program planning.

Step 3: Determine Timelines and Available Resources

- Consider the context in which the evaluation is occurring, and any factors and processes that may impact the overall timelines for the evaluation.
- Decide the time required, identify the resources needed to carry out the evaluation. These include budget available, staff involved, supplies, equipment and other tools needed for both implementing and evaluating the program.
- Example,
  Planning a 6-month timeline with budget for posters, food-demonstration materials, and one enumerator for evaluation.

Step 4: Develop Evaluation Questions

- Formulate clear questions that the evaluation must answer.
- Since, this evaluation model is a goals-based model, which can be used to measures processes or outcomes. Evaluation questions differ according to what is being measured.
- Example evaluation questions,

Did students' junk-food consumption decrease?

Were program activities implemented as planned?

Step 5: Select Measurement Methods and Procedures

- In this step, determine what to measure, and what procedures (tools and methods) to use in order to measure it.
- This includes how, when, and from whom the data will be collected, with consideration for ethical conduct related to data collection.
- Examples, Pre and post-surveys, Observation checklist in school canteen, Focus group with students

Step 6: Create the Evaluation Plan

- The evaluation plan documents all of the decisions and information produced in the model.
- Typically, an evaluation plan includes the program description, the purpose of the evaluation, evaluation questions and methodology, a data analysis plan, budget and timelines, and how the results of the evaluation will be used.

Step 7: Collect Data

- Gather data according to the tools and schedule developed.
- To ensure that data are reliable, develop standard data collection procedures and tools, and provide training for those collecting data.
- Example, Administering questionnaires to students before and after the nutrition sessions, observing canteen sales for 2 weeks.

Step 8: Process and Analyze Data

This step involves engaging with the data to process, analyze, and synthesize information from all sources.

- Ensure quality data
- Organize the data
- Analyze the data
- Synthesize the data

Step 9: Interpret and Disseminate Results

- This step involves describing what was learned through the evaluation, interpreting the data analyzed and synthesized, so that decisions can be made about the program.
- The appropriate channels/formats for the audience and purpose can be selected for dissiminating the evaluation results.
- Presentation of findings can take many forms such as a written report, slide show presentation, infographic, and/or short informational video.
- Example, “Junk-food buying decreased by 35%”

Step 10: Apply Evaluation Findings

- The ultimate purpose of program evaluation is to use the information.
- Evaluation findings can be used to modify, continue, scale up, or discontinue the program.
- Example,

  School bans packaged junk food during school hours

  Program expanded to all grades in the school

Strengths

- Very easy to implement
- Objective-based and measurable
- Good for quantitative evaluation
- Fits well with RBM log-frame, M&E plans

Limitations

- Ignores unintended effects
- Requires well-defined goals (which many programs lack)
- Can be biased if goals are unrealistic or politically influenced

Decision Making Model

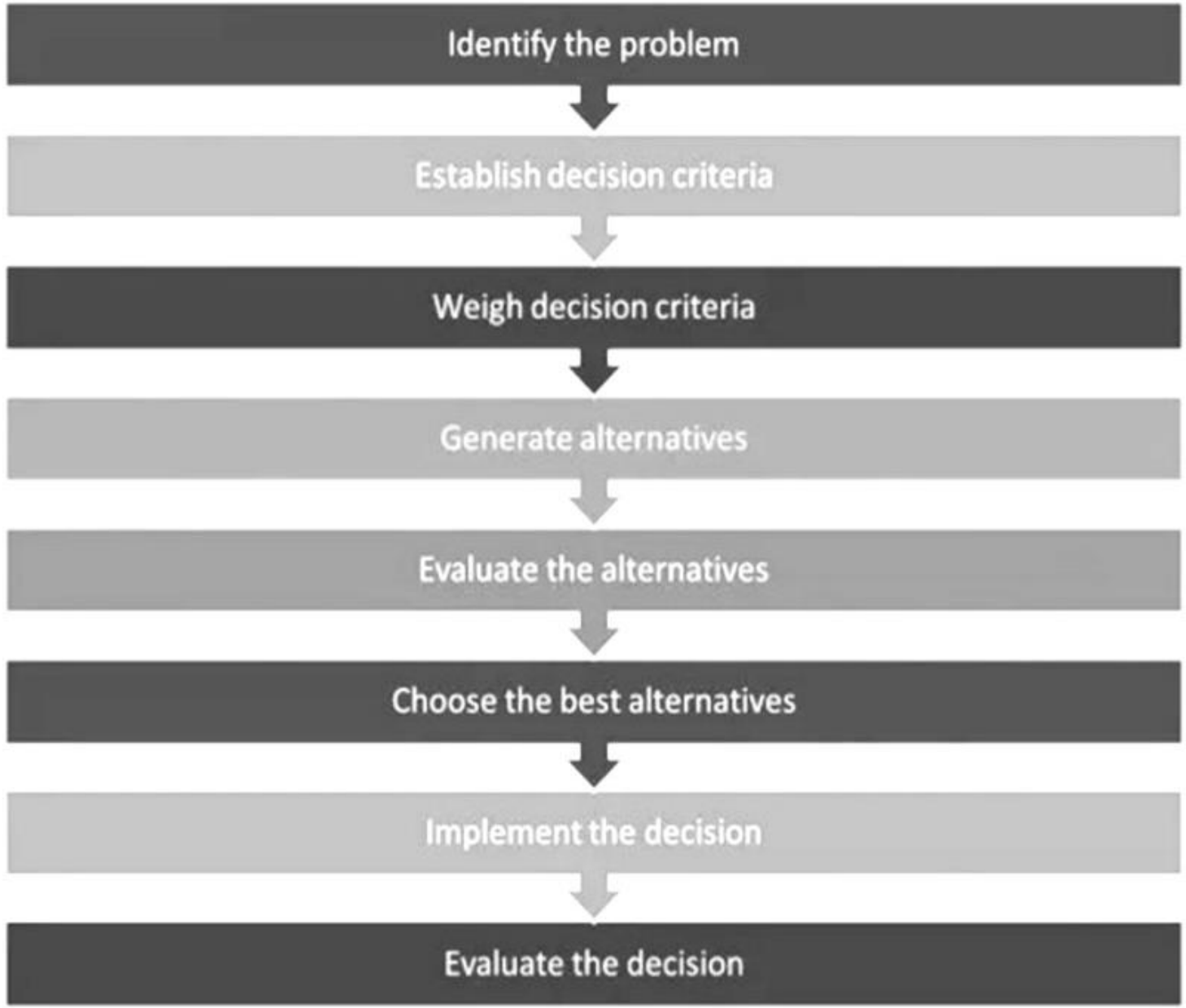
- The model was developed by Daniel Stufflebeam (also linked to CIPP philosophy)
- The key idea of decision making model is that, “Evaluation is for decision-making, not just for judging
- Decision making means the process of selecting the best choices among various options.
- The decision making process is a method of gathering information, assessing alternatives, and making a final choice with the goal of making the best decision possible.

Evaluation provides information for decisions such as:

- continue or discontinue a program
- modify program design
- allocate resources
- scale up or scale down
- introduce new strategies

Decision-making model is the rational decision- making model which consists of 8 steps that decision makers need to take to achieve the optimal decision given their goals and constraints

Rational Decision Making Model



8 Steps of the Rational Decision-Making Model

1. Identify the problem

Clearly define what issue needs a decision.

- Example, Problem: The municipality finds that full immunization coverage among children 12–23 months is only 72%, below the national target of 90%.

2. Establish decision criteria

List the factors that must be considered (cost, time, feasibility, effectiveness).

- Example, Cost, Feasibility, Community acceptance, Human resource availability, Time required, Expected impact on coverage

3. Weigh the decision criteria

Decide which criteria are more important (assign weightage).

- Example, Impact on coverage → High weight,

  Feasibility → High weight, Cost → Medium weight,

  Time → Medium weight, Community acceptance → High weight

4. Generate alternatives

List all possible solutions or options.

- Example, Conduct weekly outreach vaccination clinics, Mobilize FCHVs for door-to-door counseling, Strengthen defaulter tracking using HMIS

5. Evaluate the alternatives

Compare each alternative against the weighted criteria.

- Example, Outreach clinics → Highly effective but costlier

  FCHV mobilization → Very feasible and accepted

  Defaulter tracking → High impact, low cost

6. Choose the best alternative

Select the most suitable option based on evaluation.

- Example, Based on the weighted scores municipality chooses: Strengthening defaulter-tracking + FCHV mobilization

7. Implement the decision

Put the chosen option into action.

- Example, home visits by FCHV, Monthly defaulter-tracking meetings at the health post

8. Evaluate the decision

Monitor results and see if the problem is solved.

- Example, After 6 months Full immunization coverage increases from 72% to 89%

Strengths

- Practical and action-oriented, Directly useful for policymakers
- Helps in resource allocation
- Visualize the decisions and rules
- Facilitate collaboration

Limitations

- Time consuming
- Limited analysis
- Evaluation becomes political and chances of biased decisions
- Decision-makers may ignore results
- Focus may shift from improvement to justification

Goal Free Evaluation Model

- Goal free evaluation model was developed by Michael Scriven (British born Australian philosopher) in 1972.
- Goal free evaluation is any evaluation in which the evaluator conducts the evaluation of a program without knowing its goals to avoid bias.
- The Goal free evaluator attempts to observe and measure all actual outcomes, effects or impacts, intended or unintended , all without being cued to the program's intentions.
- GFE evaluator asks what does the program actually do? Rather , what does the program intend to do?

When to use goal free Evaluations?

- When stakeholder want information about program outcomes ,both intended and unintended
- When Evaluator have no knowledge of program goals , intentionally or unintentionally
- Want to identify the effect of a program from data collection , observations and interviews.

Methodologies

- Determine what effects this program had and evaluate whether or not they were intended
- Evaluate the actual effects against a profile of demonstrated needs
- Determine if what occurred can logically be attributed to the program or intervention
- Determine the degree to which the effect is positive, negative or neutral
- Do not be under the control of the Management, choose the variables of the evaluation independently
- The key to goal free evaluation is to have an evaluator enter the field and try to learn about a program and its results with out being aware of the objectives of program.
- Supplement to the more traditional goal oriented evaluation

Implementation Technique for GFE

- Reviewing expert opinion
- Literature review
- Visiting sites
- Examine similar program
- Ask questions
- Determine what the evaluation will do & how it will benefit to stakeholders
- Select appropriate methodology for gathering data
- Identify key and critical issues
- Unexpected outcome
- Draw out key issues
- Provide appropriate feedback

Strengths of GFE

- Controls goal manipulation and political influence
- Limits bias & broadens the area of evaluation when goal is not concerned
- Detects unplanned/unexpected outcomes
- Aligning goals with actual program activities and outcome
- Useful when goals are vague or exaggerated

Criticism

- Evaluator may miss important intended outcomes
- Not suitable for large, multi-component programs
- Hard to collect data without framework
- This approach only can lead to poor planning

Model	Focus	Strength	Limitation
System Analysis	Inputs–Processes–Outputs–Feedback	Simple, structured	Doesn't explain "why"
Goal Attainment / Behavioral Objectives	Achievement of goals	Easy, measurable	Ignores unintended outcomes
Decision-Making	Information for decisions	Useful for managers	May be politically influenced
Goal-Free	Actual outcomes, ignoring goals	Unbiased, finds hidden effects	Hard to guide evaluation

Evaluation Design

- Evaluation design is the structured plan or blueprint for how an evaluation will be conducted.
- It outlines **what will be measured** , **how data will be collected**, **when it be collected, who will be involved, and what methods will be used** to determine the effectiveness, impact, or value of a program ,project or intervention.
- Evaluation design is not the same as the 'research methods' but it does help to clarify which research methods are best suited to gathering the information (data) needed to answer the evaluation questions.

Important points to consider when deciding on an evaluation design are:

- The questions you want to answer
- The audience for the evaluation
- The maturity of your program (i.e. is it ready to evaluate outcomes or has it only just started?)
- The type of program or intervention you are seeking to evaluate
- Your client or target group (e.g. who the program is for, how many people are in the program or receive a service and what their characteristics are)
- What data are already available
- Your resources (e.g. funding, staff, skills) and time frame
- Whether you will conduct an evaluation internally or contract an external evaluator.

Types of Evaluation Designs:

1. Internal and External evaluation trade off
2. Formative and summative evaluation design
3. Retrospective and prospective evaluation design
4. Interventional and non-interventional evaluation design
5. Experimental design

a. Pre- experimental
b. Quasi- experimental
c. True experimental

6. Non-experimental designs

1. Internal and external evaluation tradeoffs

•**Internal evaluation (self evaluation):** in which people within a program sponsor, conduct and control the evaluation.

For example: a school teachers assesses their own teaching methods and students' performance.

•**External evaluation:** in which someone from beyond the program acts as the evaluator and controls the evaluation.

For example: an independent consultants evaluates a school's teaching quality and performance.

Trades off between Internal and External Evaluation:

Aspect / Point	Internal Evaluation	External Evaluation
1. Cost	Lower, uses existing staff/resources	Higher, requires external funding or contracts
2.Organizational Knowledge	Deep understanding of program history, culture, and processes	Limited inside knowledge but broader external perspective
3. Flexibility	Easily adapted to program timelines and needs	Less flexible, bound by external schedules/contracts
4. Bias Risk	Higher risk of subjectivity due to vested interests	Lower risk, more objective and independent
5. Expertise	May lack specialized evaluation skills	Brings advanced skills and wider methodological experience
6. Credibility	Findings may be questioned by funders or external stakeholders	Results carry more weight with donors, policymakers, and the public
7. Control	Organization has direct control over evaluation design and process	Less control, external evaluators set methods and scope
8. Program Improvement	Stronger focus on internal learning and continuous improvement	Focused on accountability, compliance, and external reporting
9. Stakeholder Engagement	Easier to involve staff and beneficiaries in evaluation	May struggle to build trust with program participants
10. Sustainability	Builds long-term internal capacity for ongoing program evaluation	Provides one-time expertise but less sustainable internally

2. Formative vs. Summative designs:

Formative evaluation:

- Formative evaluation **(formative assessment)** is a process used to **monitor a program or performance during the development of a program, course or skill** with the primary goal of **improving and guiding future learning of development**.
- Its goal is to monitor progress , identify strengths, weaknesses and make necessary adjustments to enhance effectiveness before a final product or outcome is reached.

- In formative evaluation, programs are typically assessed during their development or early implementation to provide information about how best to revise and modify for improvement.
- This type of evaluation often is helpful **for pilot projects and new programs,** but can be used for **progress monitoring of ongoing program**.

Summative evaluation:

- Summative evaluation (summative assessment) is an evaluation that occurs **at the end of a learning period, program, or project to measure the overall effectiveness, achievement, or outcomes.**
- It is used to judge the success of a program against predetermined goals often resulting in a grade or a decision about the program's continuation.
- Its goal is **to evaluate student learning at the end of an instructional unit by comparing it against some standard or benchmark**.
- Example of summative evaluation include: **a midterm exam**

Formative Evaluation Vs Summative Evaluation

Points	Formative Evaluation	Summative Evaluation
1. Purpose	To improve ongoing teaching and learning	To judge overall learning at the end of instruction
2. Timing	Conducted during the learning process	Conducted after a unit, term, or course is completed
3. Focus	Identifies strengths, weaknesses, and learning gaps	Measures achievement of learning objectives
4. Feedback	Frequent, immediate, and descriptive	Final, often in the form of scores or grades
5. Use	Helps teachers adjust instruction and students improve	Used for certification, grading, and accountability
6. Examples	Quizzes, class discussions, drafts, observations	Final exams, projects, standardized tests
7. Stakes	Low-stakes; supports learning	High-stakes; evaluates learning
8. Nature	Continuous and ongoing	Final and conclusive

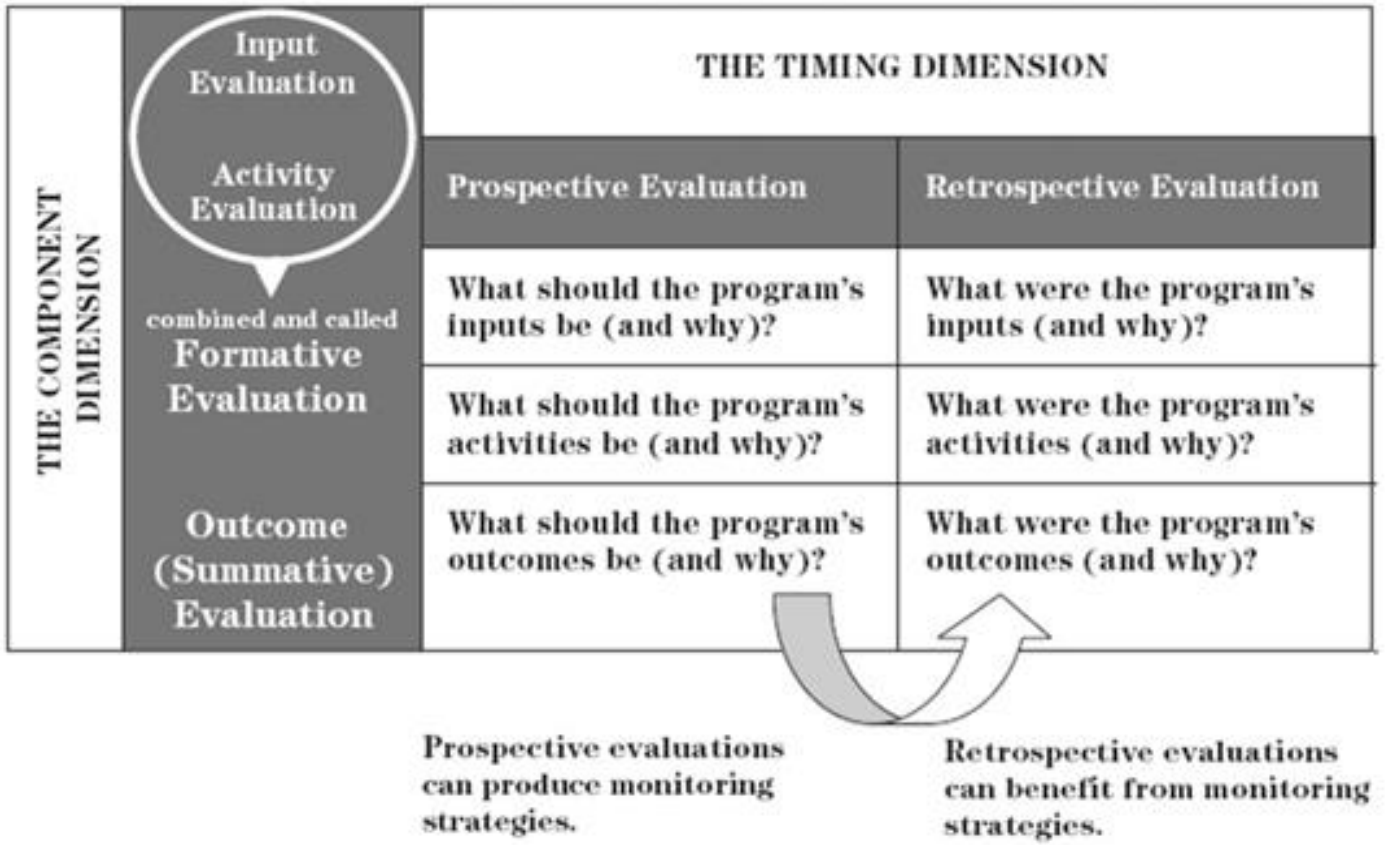
3. Retrospective and Prospective design:

Retrospective Evaluation:

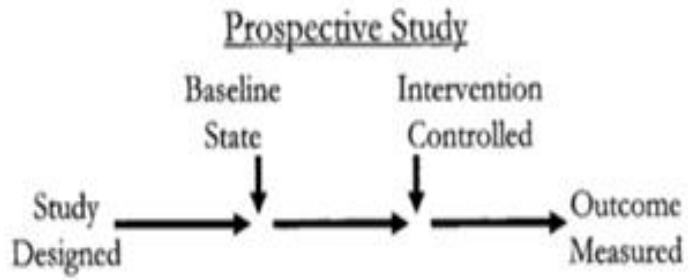
- Retrospective evaluation is a type of assessment conducted after **project, program, process, or event has been completed.**
- It' s purpose is to **look back and analyze what happened, why it happened ,and how outcomes compare to goals or expectations.**
- However, evaluation should also be conducted when a new program within a service is being introduced.

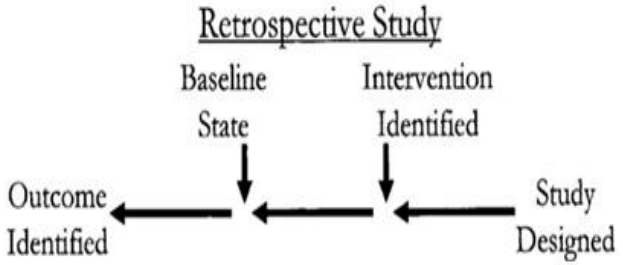
Prospective Evaluation:

- Prospective evaluation is a type of assessment conducted before or during the **early stages of a project program or intervention.**
- It identifies ways to increase the impact of a program on clients; it examines and describes a program's attributes; and, it identifies **how to improve delivery mechanisms** to be more efficient ,effective ,and responsive to clients needs.
- It' s purpose is **to predict, plan, and guide future actions rather than analyze past performance.**


THE COMPONENT DIMENSION
Input Evaluation
Activity Evaluation
combined and called Formative Evaluation
Outcome (Summative) Evaluation
THE TIMING DIMENSION
Prospective Evaluation
Retrospective Evaluation
What should the program's inputs be (and why)?
What were the program's inputs (and why)?
What should the program's activities be (and why)?
What were the program's activities (and why)?
What should the program's outcomes be (and why)?
What were the program's outcomes (and why)?
Prospective evaluations can produce monitoring strategies.
Retrospective evaluations can benefit from monitoring strategies.

Design of Prospective study





Retrospective vs. Prospective Evaluation:

Point	Retrospective Evaluation	Prospective Evaluation
1. Timing	Conducted *after* an intervention or event has occurred.	Conducted *before or during* an intervention or event.
2. Data Source	Uses existing or historical data.	Collects new, forward-looking data.
3. Control Over Variables	Limited ability to control confounding factors.	Greater control over variables through planned study design.
4. Bias Risk	Higher risk of recall, selection, and information bias.	Lower bias because data is collected in real time with predefined methods.
5. Cost & Resources	Generally cheaper and quicker; uses available data.	More expensive and time-consuming due to active data collection.
6. Causality Assessment	Weaker at establishing causation; mainly identifies associations.	Stronger ability to establish causal relationships.
7. Flexibility	Flexible in exploring unanticipated outcomes in existing records.	Less flexible; design must be set in advance.
8. Ethical Considerations	Fewer ethical issues since no interference with participants.	Requires approvals and ethical oversight for data collection.
9. Typical Use Cases	Audits, historical analysis, quality reviews, chart studies.	Clinical trials, cohort studies, future risk prediction, program monitoring.

4. Interventional vs. Non interventional design:

Interventional evaluation:

- Interventional evaluation is the systematic assessment of a program to determine its effectiveness, outcomes, and impact by comparing results to a baseline and understanding how, why, and for whom it works.
- It provide to seek reliable evidence for decision making ensuring resources are used well, identifying strengths, correcting flaws and potentially preventing diseases, treating existing conditions or restoring function.
- This process is used in various fields like public health, clinical research, and project management to make informed judgements about an intervention's success, identify gap between planned and actual results and improve future performance.

Non interventional evaluation:

- A non interventional evaluation is also known as a non interventional study(NIS) is an observational study of how a medicine or medical device is used in routine clinical practice.
- Unlike an interventional study (Clinical trial), it doesn't require additional diagnostic or monitoring procedures beyond those of normal practice, nor does the study protocol dictate the treatment a patient receives.
- Its goal is to collect real-world data on safety, effectiveness, and other aspects of a product's use in a diverse patient population under everyday conditions.

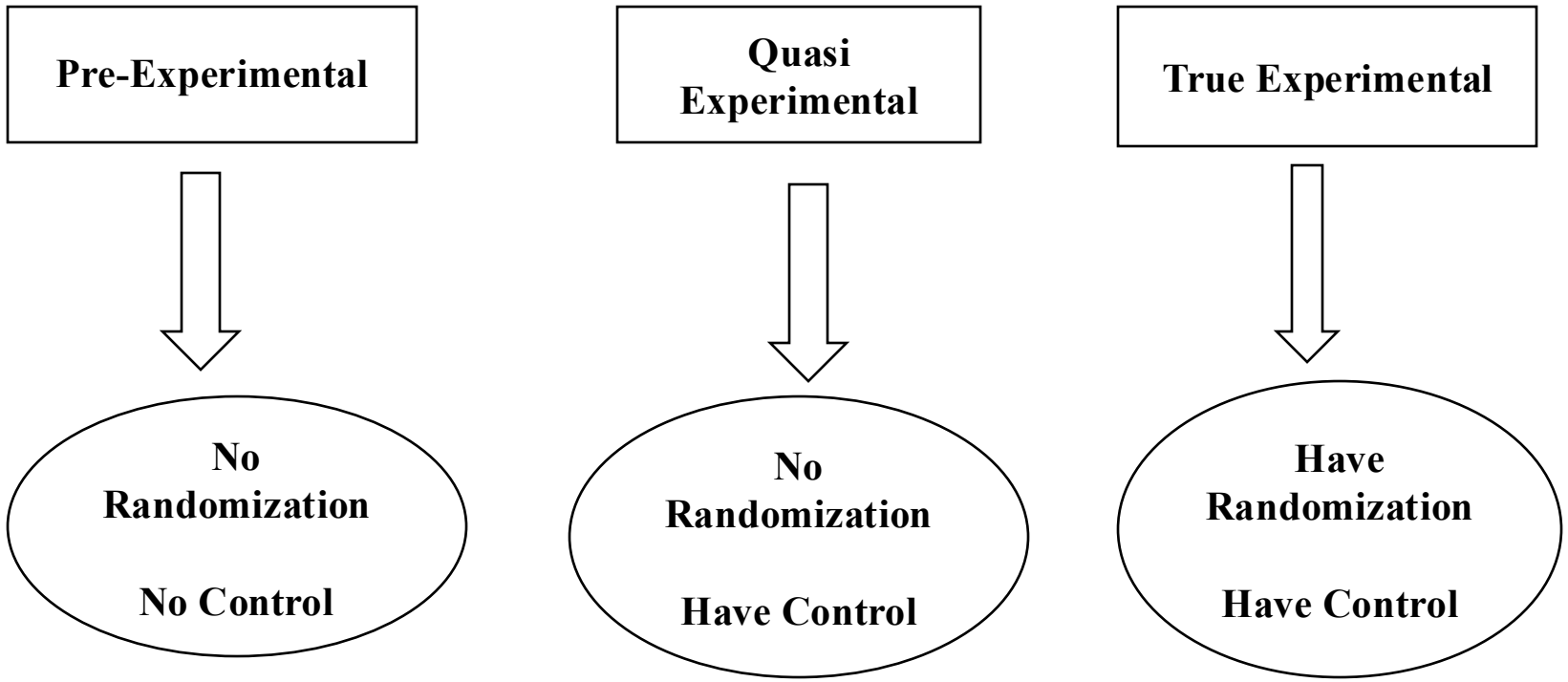
Interventional vs Non - interventional evaluation:

Point	Interventional Evaluation	Non-Interventional Evaluation
1. Definition	Researcher actively assigns treatment or intervention	Researcher only observes routine care without interference
2. Purpose	Test efficacy and safety of new drugs/procedures	Collect real-world evidence of existing treatments
3. Control of Variables	Controlled environment, often randomized	Natural setting, no control over variables
4. Example	Clinical trial testing a new cancer drug vs placebo	Observing diabetic patients using insulin in daily life
5. Ethical Considerations	High ethical risk, requires strict approval	Lower ethical risk, minimal intervention
6. Data Collection	Structured trial protocols and monitoring	Medical records, registries, patient surveys
7. Cost & Complexity	Expensive, complex, requires infrastructure	Cheaper, simpler, often uses existing data
8. Regulatory Oversight	Strict clinical trial regulations	Fewer regulatory requirements, but ethical standards apply
9. Outcome Reliability	Strong internal validity (controlled design)	Strong external validity (real-world practice)
10. Use in Healthcare	Approving new drugs, devices, therapies	Monitoring long-term safety, effectiveness, adherence

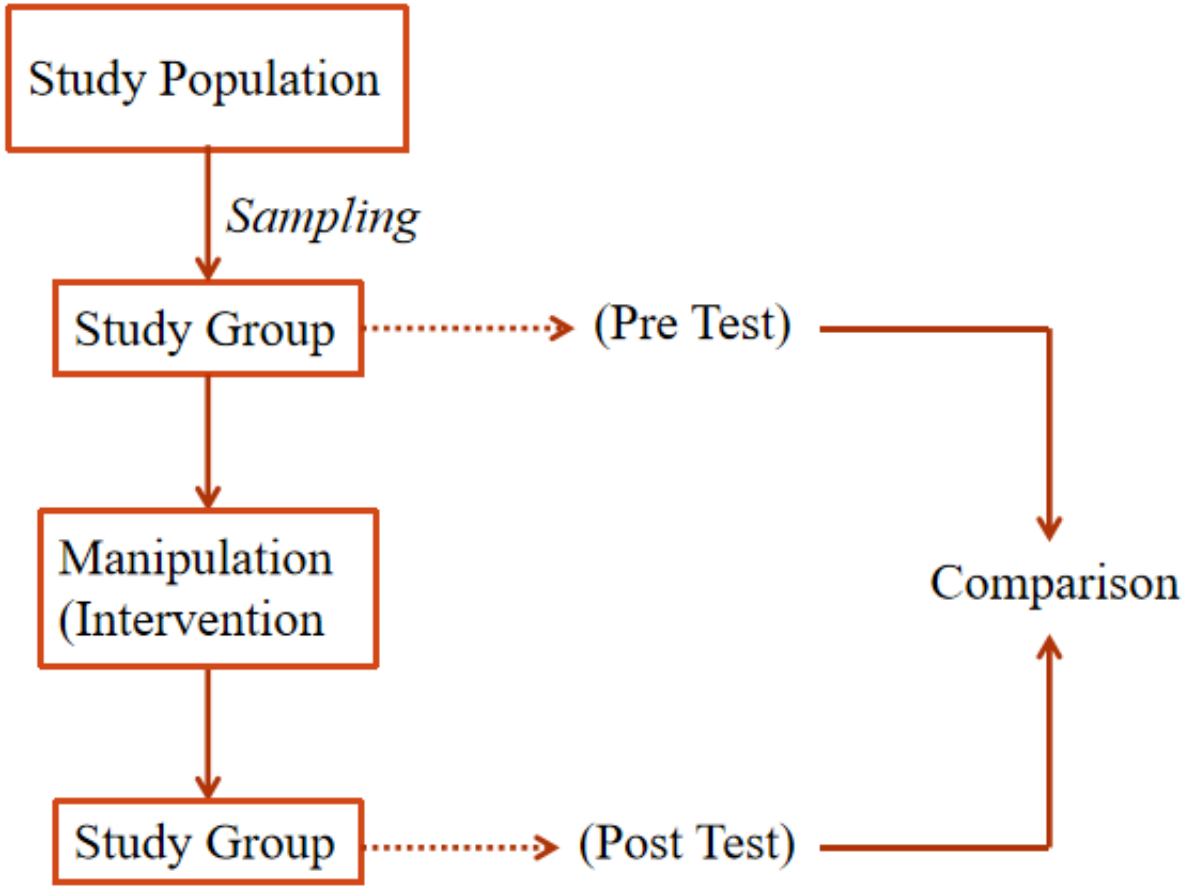
5. Experimental/ interventional design

- Design in which subject are randomly assigned from a single population to the experimental group and control group.
- **Treatment group/interventional group**: who receives treatment and effects are studied (e.g. new medication)
- **Control group**: who do not receive treatment/intervention(e.g. Placebo)
- Used to determine if a program or intervention is more effective on participant's health outcomes, behaviors and knowledge.

Types of Experimental/ interventional designs

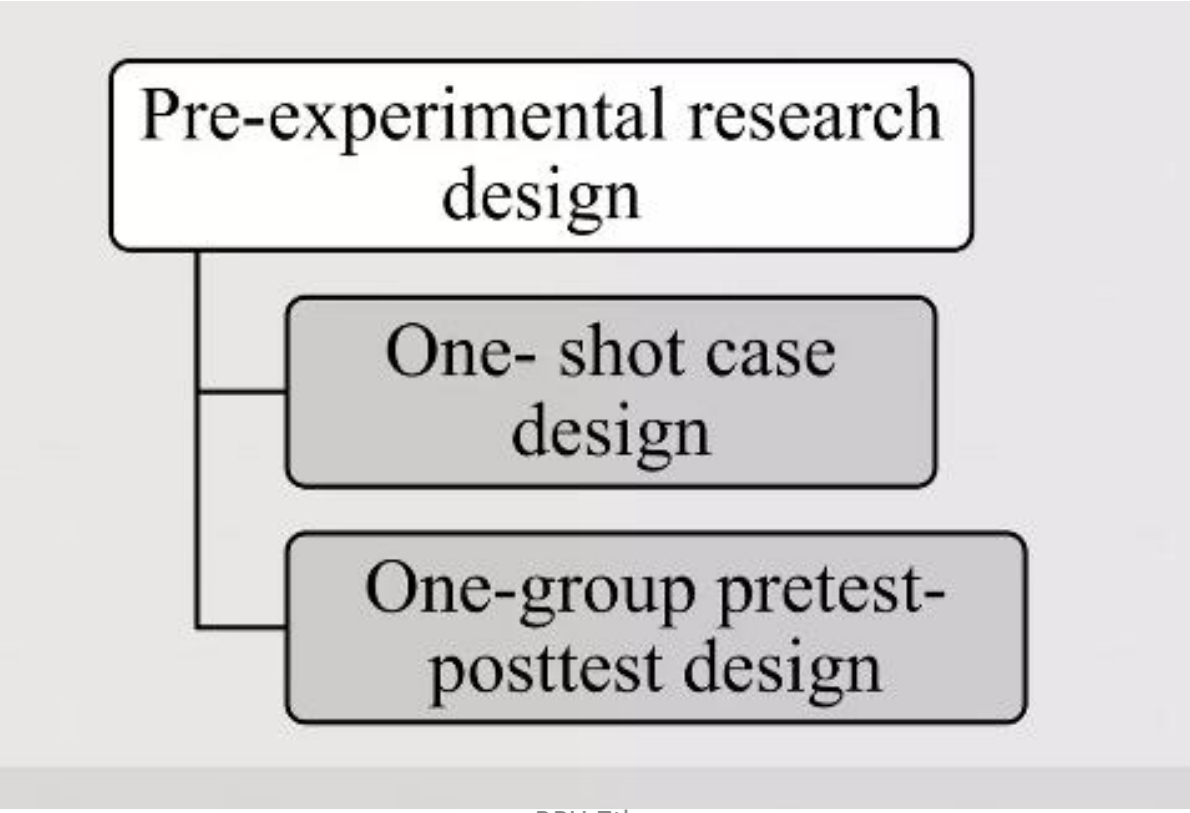


Pre-Experimental Research Design



Pre-experimental design types

- This design is considered very **weak**, because the researcher has very little control over the experiment.

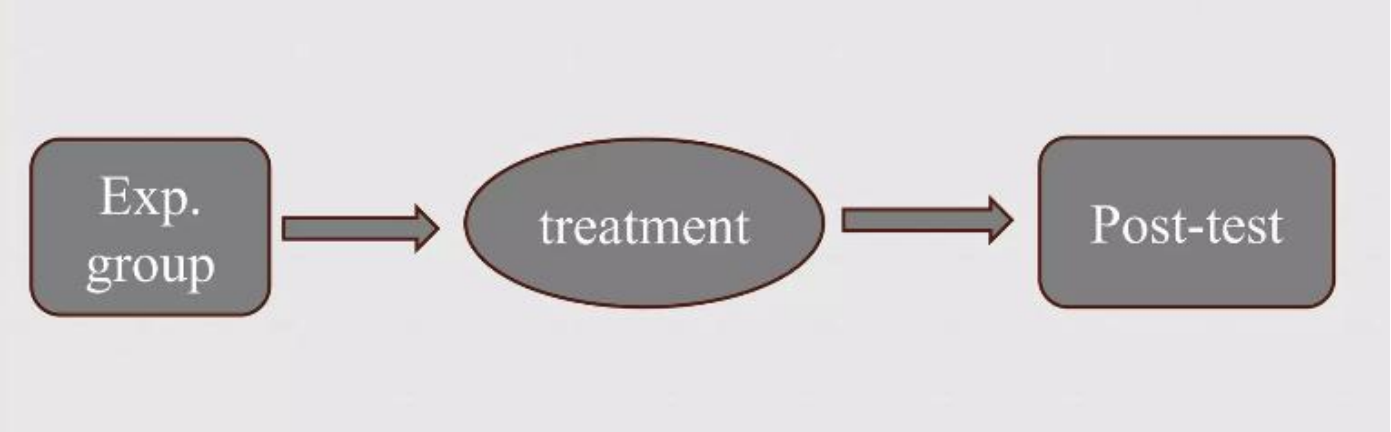


One shot case study

- In this research design, a **single experimental group** is exposed to a treatment & observations are made after the implementation of that treatment.
- There is **no random assignment** of subjects to the experimental group & **no control group** at all.

One shot case study

Example: suppose we wish to see if a new textbook increases students' interest in the course(history, science etc)

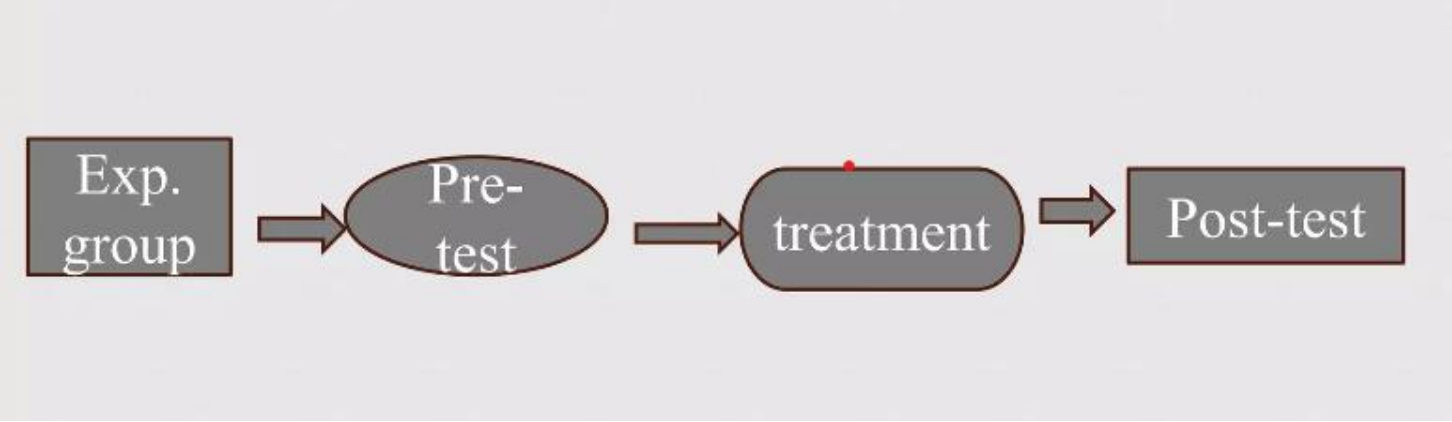


One group pre-test post-test design

- A type of pre-experimental design where pre-test and post-test are taken with intervention but fails to include randomization and control.
- It is the simplest type of pre-experimental design, where only the experimental group is selected as the study subjects.
- A pretest observation of the dependent variables is made before implementation of the treatment to the selected group; the treatment is administered & finally a posttest observation of dependent variables is carried out to assess the effect of treatment on the group.

One group pre-test post-test design

Example: suppose we want to assess the effects of weekly counseling sessions on the attitudes of identified bullies in school.



Pre-experimental design

Advantages

- Very simple & convenient to conduct these studies in natural settings, especially in nursing.
- Most suitable design for the beginners in the field of experimental research.

Disadvantage

- Considered a very weak experimental design to establish causal relationship between independent & dependent variables, because it controls no threat to internal validity. It has very little control over the research.

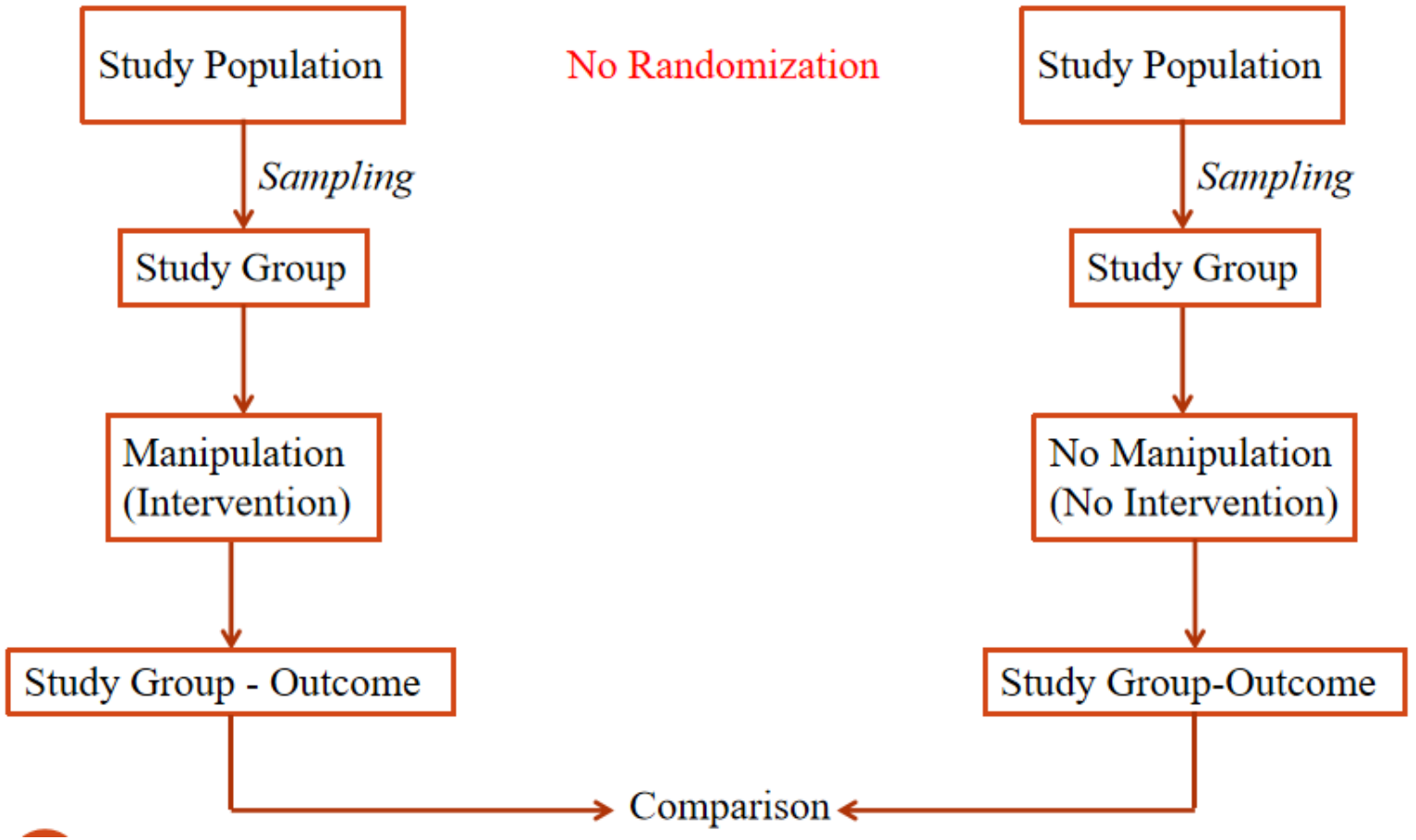
Quasi- experimental design

- “Quasi” Latin word means **almost but not really**.
- Quasi-experimental research design involves the **manipulation** of independent variable to observe the effect on dependent variable, but it lacks at least of the two characteristics of the true experimental design; randomization or a control group.

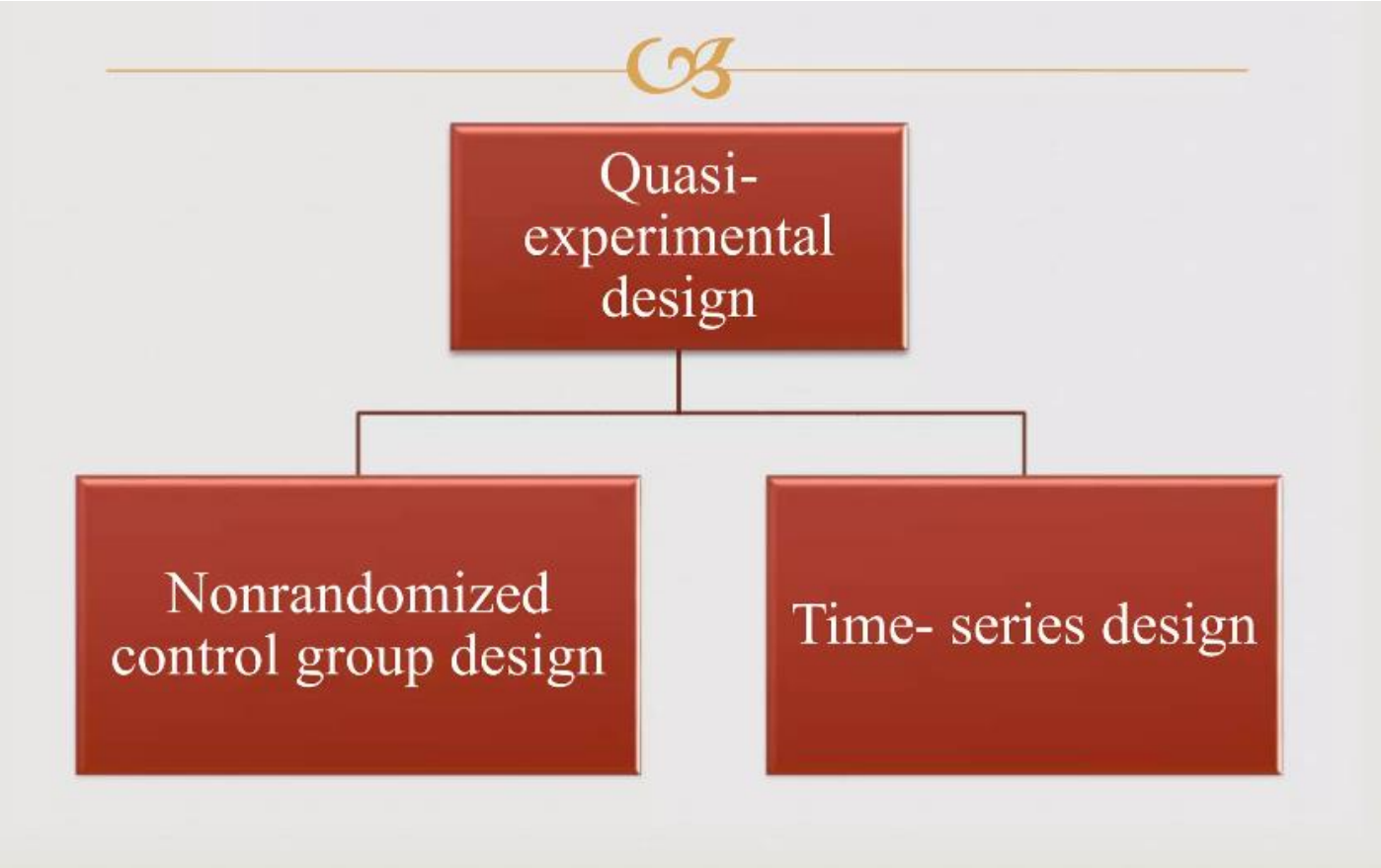
Main characteristics

- **Manipulation** of the independent variables to observe the effects on the dependent variables.
- Lack of at least one of the two other essential characteristics of the true experiment. i.e. random assignment of subject or a control group.
- Quasi-independent variables are used instead of true independent variables. Where independent variable is not manipulated in complete controller situations.

Quasi-Experimental Research Design



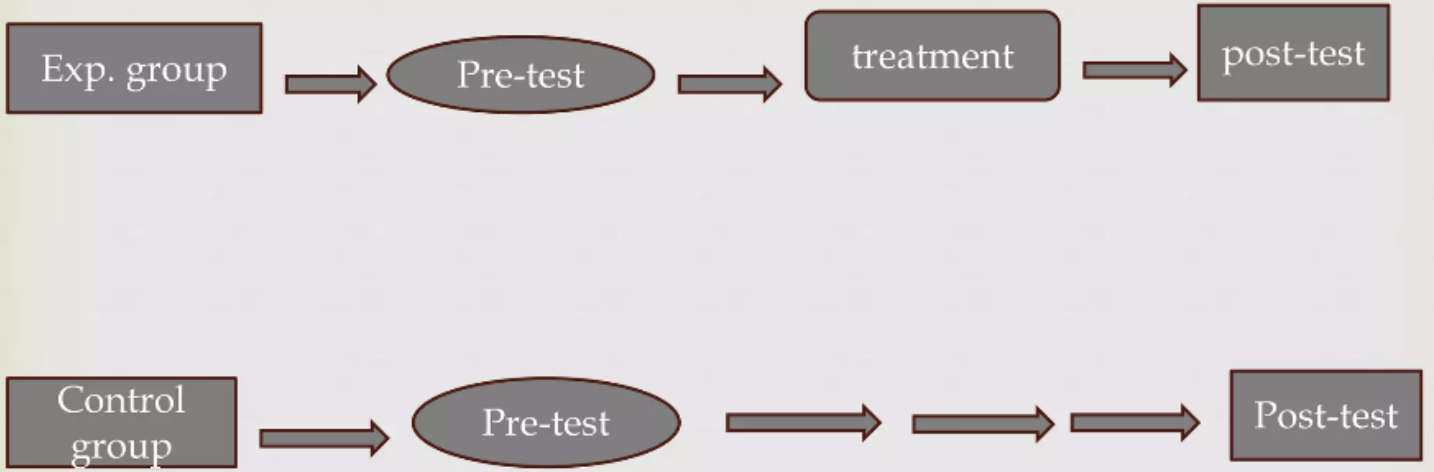
Types of quasi-experimental design



Non-randomized control group design

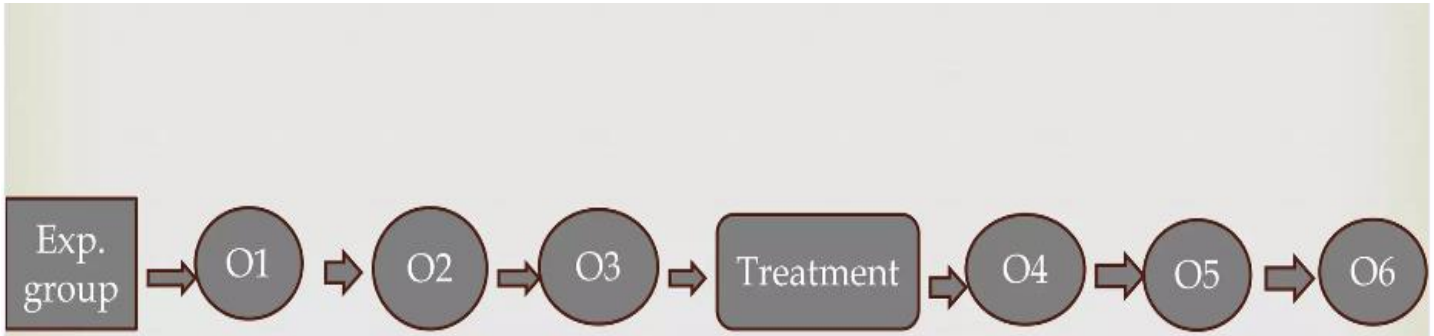
- It is also known as the “nonequivalent control group design”.
- This design is identical to the pretest-posttest control group design, except there is no random assignment of subjects in experimental & control groups.
- In this design, experimental & control groups are selected without randomization, & dependent variables are observed in experimental as well as control groups before the intervention
- Later, the experimental group receives treatment & after that posttest observation of dependent variables is carried out for both the groups to assess the effects of treatment on experiment group.

Non randomized control group design



Time series design

- This design is useful when the experiment wants to measure the effects of a treatment over a **long period of time**.
- The experiment would continue to administer the treatment & measure the effects a number of times during the course of the experiment.
- Generally, it is single-subject research, in which the researcher carries out an experiment on an **individual** or on a **small number of individuals**, by alternating between administering & then withdrawing the treatment to determine the effectiveness of the intervention.



If no change between O1 and O2 or O2 and O3 but a significant difference between O3 and O4 will be found ,only then it can be said that the treatment is effective.

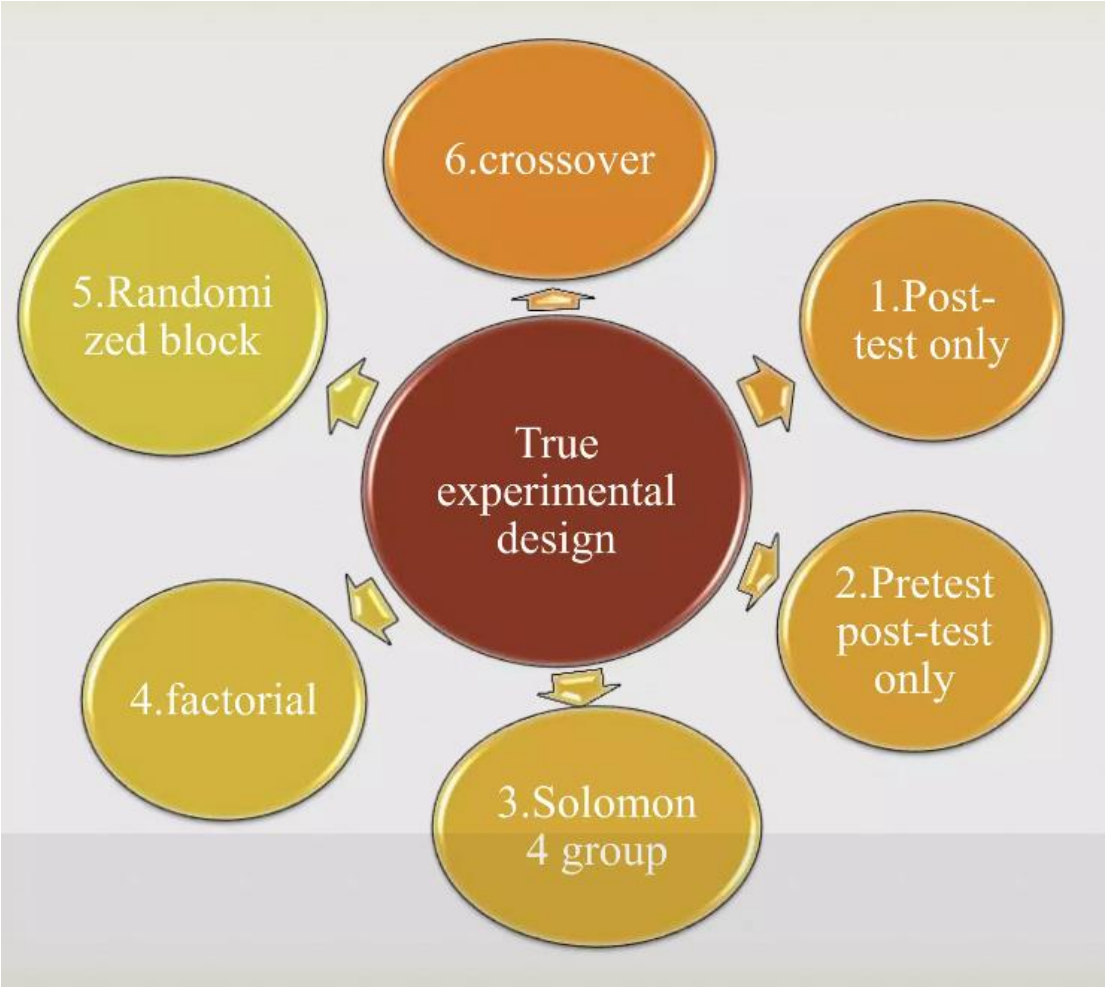
Advantages of quasi-experimental design

- Quasi- experimental designs are **more frequently used** because they are more practical & feasible to conduct research studies in different field, where in the absence of a large sample size, randomization &/ or availability of control groups are not always possible.
- This design is more suitable for **real-world natural setting** than true experimental research designs.
- It allows researchers to **evaluate** the impact of quasi-independent variables under naturally occurring conditions.
- It may be able to **establish causal relationship**. Wherein some of hypothesis are practically answered through this design only.

True- experimental design

- The true experimental research design relies on statistical analysis to **approve or disprove a hypothesis**. It is the most **accurate** type of experimental design and may be carried out with or without a pretest on at least 2 randomly assigned dependent subjects.
- The true experimental research design must contain a control group, a variable that can be manipulated by the researcher, and the distribution must be random.

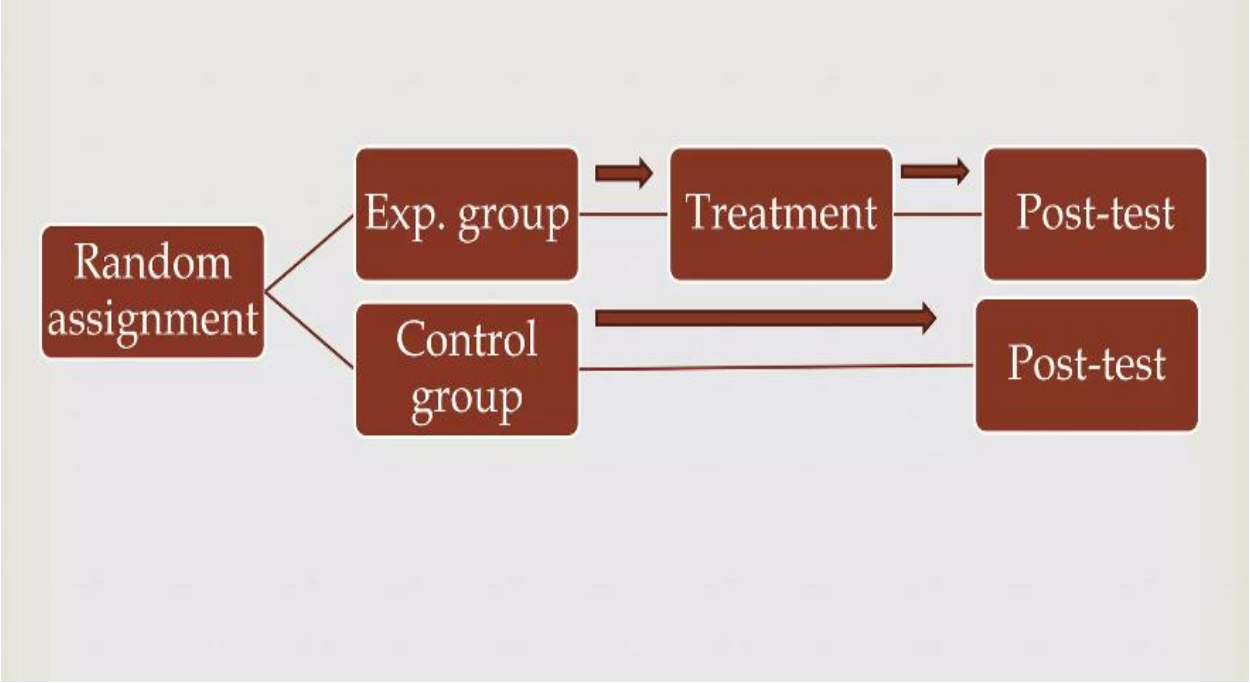
Types of True- experimental design



Post- test only group design

- Composed of two randomly assigned group, i.e. experimental and control, but **neither** of which is pretested before the implementation of treatment on the experimental group.
- In addition, while the treatment is implemented on the experimental group only, post-test observation is carried out on both the group to access the effect of manipulation.

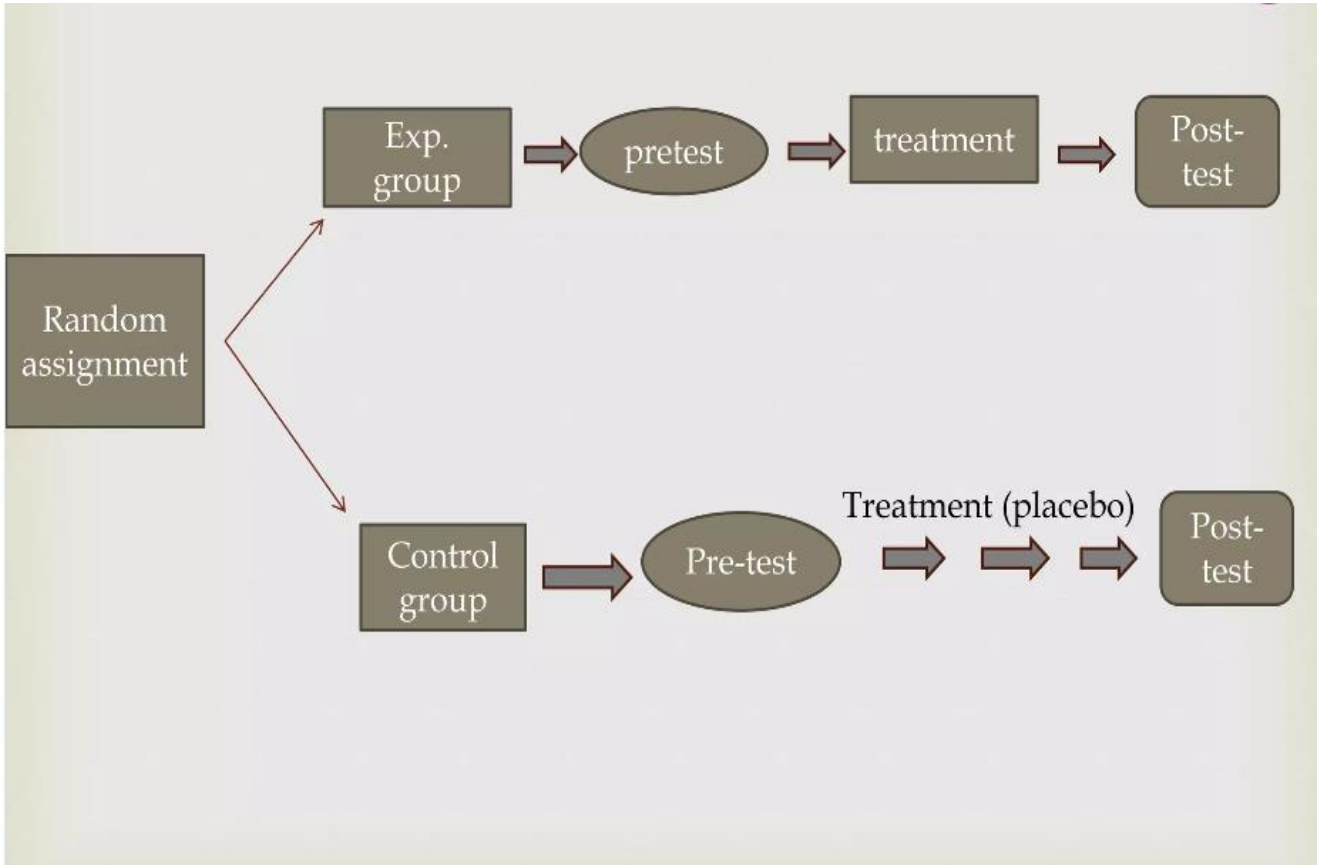
Post- test only group design



Pre test post test only design

- In these experiment the researcher conducts experimental group and control group.
- The effect of the dependent variable on both the groups is seen before the treatment (pretest).
- Later the treatment is carried out on experimental group only, & after treatment observation of dependent variable is made on both the groups to examine the effect of the manipulation of independent variable on dependent variable.
- For example, such a design could be used for an experimental study to assess the effectiveness of cognitive behavioral therapy interventions for patients with breast cancer.

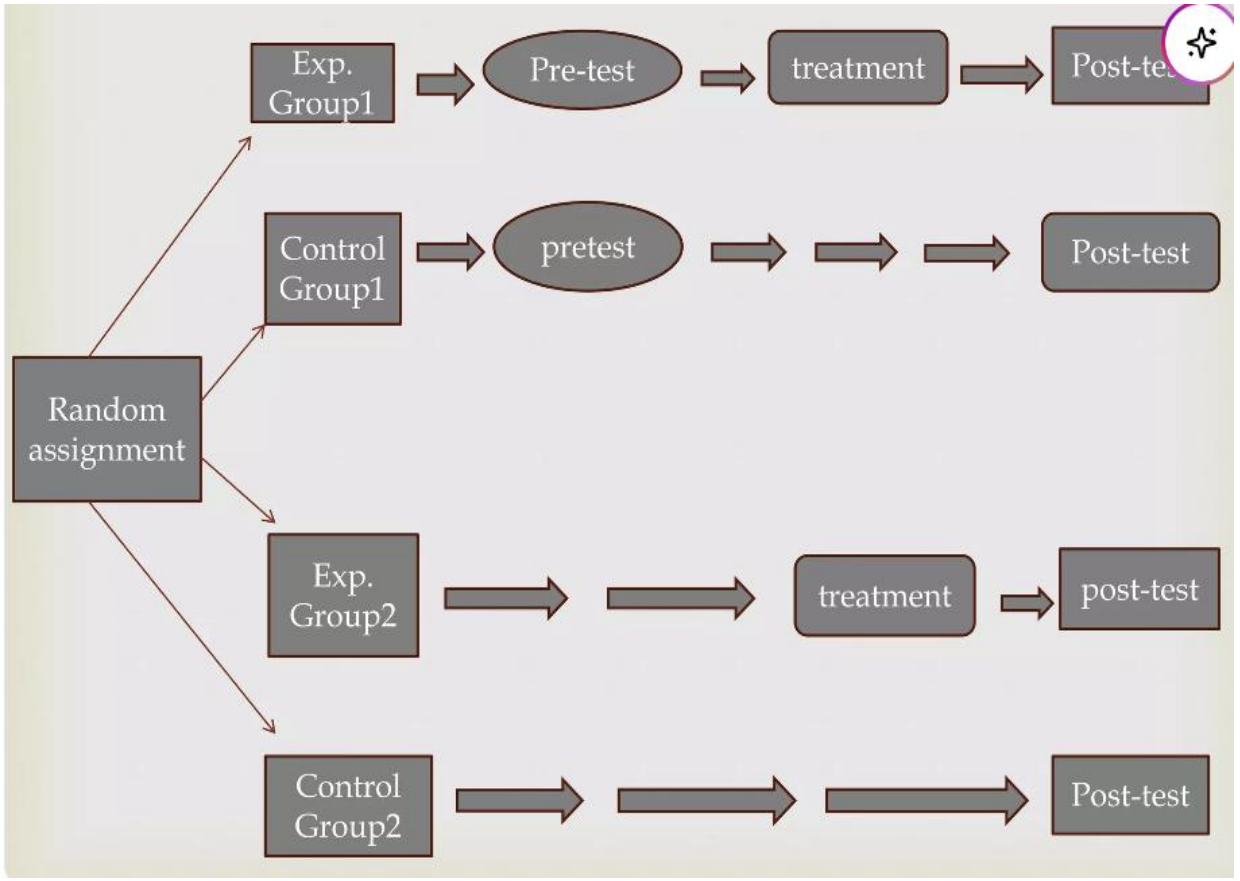
Pre test post test only design



Solomon 4 group design

- There are two experimental groups(experimental group 1 & experimental group 2) & two control group (control group1 & control group 2)
- Initially, the investigator randomly assigns subjects to the four groups.
- Out of the four groups, only experimental group 1 & control group 1 receives the pretest, followed by the treatment to the experimental group 1 & experimental group 2.
- Finally, all the 4 groups receive post-test, where the effects of the dependent variables of the study are observed & comparison is made of the 4 groups to assess the effect of independent variable(experimental treatment) on the dependent variable.

Solomon 4 group design



Factorial design

- In factorial design, researcher manipulates two or more independent variables simultaneously to observe their effects on the dependent variables. This design is useful when there are more than two independent variables, called factors to be tested.
- This design also facilitates the testing of several hypothesis at a single time.
- Typical factorial design incorporates 2*2 or 2*3 factorial, but it can be in any combination.

Factorial design

Frequency of mouth care	Protocols of mouth care	
	Chlorhexidine($\alpha 1$)	Saline($\alpha 2$)
4 hourly($\beta 1$)	$\alpha 1 \ldots . \beta 1$	$\alpha 2 \ldots . \beta 1$
6 hourly($\beta 2$)	$\alpha 1 \ldots . \beta 2$	$\alpha 2 \ldots . \beta 2$
8 hourly($\beta 3$)	$\alpha 1 \ldots . \beta 3$	$\alpha 2 \ldots . . \beta 3$

The first number (α) refers to the independent variables or the type of experimental treatments, & the second number (β) refers to the level or frequency of the treatment.

Randomized block design

- Control of inherent differences between experimental subjects & differences in experimental conditions is one of the difficult problems faced by researcher in biological sciences.
- When there are a large number of experimental comparison groups, the randomized block design is used to bring homogeneity among selected different groups.
- For example, a researcher wants to examine the effects to three different antihypertensive drugs on patients with hypertension.

In this example, to ensure the homogeneity among the subjects under treatment, researcher randomly places the subjects in homogenous groups(blocks) like patients with primary hypertension, diabetes patients with hypertension & renal patients with hypertension.

Types of antihypertensive drugs	Blocks		
	Patient with primary hypertension(I)	Diabetic patient with hypertension(II)	Renal patient with hypertension(III)
A	A,I	A,II	A,III
B	B,I	B,II	B,III
C	C,I	C,II	C,III

Crossover design

- In this design, subjects are exposed to more than one treatment, where subjects are randomly assigned to different orders of treatment.
- It is also known as repeated measures design.
- For example, when we are comparing the effectiveness of the chlorhexidine mouth care protocol on group I & saline mouth care protocol on the subjects of group II.
- Later, the treatment is exchanged where group I receives the saline mouth care & group II receives chlorhexidine. In such studies, subjects serve as their own control.

groups	Protocols of mouth care	
Group I	Chlorhexidine (α1)	Saline (α2)
Group II	Saline(α2)	Chlorhexidine(α1)

Advantages of true experimental design

- Experimental research designs are considered the **most powerful** designs to establish the casual relationship between independent & dependent variables.
- Where the purpose of research is explanation, **casual relationship** may be established among the variables by **experimentation**.
- When a true experiment is done in a lab or controlled setting, it avoids real-life distractions. This lets the researcher work more carefully, calmly and with full focus.

Comparison

Pre experimental design	True experimental design	Quasi experimental design
Least effective design	Most effective design	Less satisfactory design
No control group	Control group are present	Control group are present
Randomization may be possible	randomization	Used when randomization is not possible.

Non-Experimental /observational Design

- These designs are most appropriate for collecting **descriptive** information or for doing small case studies of a particular situation without making changes or introducing treatments.
- They are **not recommended for evaluation studies** that attempt to determine the effect of a program intervention, but they may be **useful in diagnostic** studies to determine the reasons why a problem exists.
- In this method of evaluation, only people who are participating in the program get the pre and post test.

1) Does not involve a comparison group.

2) Non-experimental designs may include pre- and post-intervention studies with no control or comparison group, case study approaches, and post-intervention-only approaches, among others.

3) The key feature of a non-experimental design is the lack of a control group.

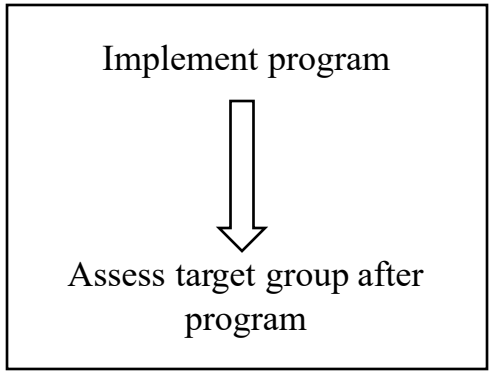
I. Single group post-test (one group, post test only)

Examines program beneficiaries after they receive program.

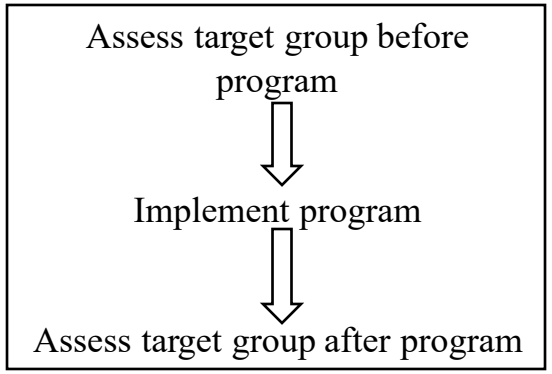
II. Single group pre and post-test (one group, pre and post test both):

Provides a comparison of program beneficiaries before and after they receive program services.

(I)



(II)



Non-experimental designs

Most of the other evaluation designs fall under the broad heading of 'non-experimental' designs. When the use of control or comparison groups is not feasible, non-experimental designs can be appropriate.

Some common non-experimental designs are:

- pre- and post-test studies
- case studies & case series
- Descriptive design
- Analytical design

Some common non-experimental designs are:

- Cross- sectional
- Longitudinal
- Prospective design
- Retrospective design
- Case- control studies
- Cohort studies

Non-experimental evaluation design

Advantages

- Fairly easy to conduct
- Requires low effort i.e. Evaluators need only hand out surveys, collect and study the data.
- Naturalistic observation; representative of real life
- Non-invasive; allows applied research when experiments manipulation isn't possible or ethical.
- The findings apply well to the real world.

Disadvantages

- Limited in establishing causal relationships due to lack of manipulation, control and randomization.
- Can fail to provide enough data to establish correlation let alone causation.
- The groups are not representative of the entire population.

Interventional vs. Non-Interventional Designs

- Interventional study designs also called experimental study designs are those where the researcher or the evaluator intervenes at some point throughout the study or project.
- The most common and strongest interventional study design is a randomized controlled trial, however, there are other interventional study designs, including pre-post study design, non-randomized controlled trials, and quasi-experiments.
- Experimental study can be used to evaluate therapeutic agents can include prophylactic agents (vaccines or drugs), treatments, surgical approaches or diagnostic tests.

Selecting an Evaluation Design

Selecting an Evaluation design depends on:

1. The Purpose of the Evaluation

- While evaluation often focuses on answering - Did the program work as planned?—it can also have other, more specific goals. An evaluation might only focus on:
  - Counting Participants [how many people were served?]
- Because these goals are so different, they require completely different plans (evaluation designs) and need people with different skills to carry them out. Therefore, clearly knowing why you are evaluating a program is the crucial first step in deciding how to evaluate it.

2. Program's Nature for Evaluation

The type of evaluation design you choose often depends on how your program is set up and how participants are involved:

- If you work with people in groups (like a class or support group), it's often easier to use designs that involve multiple groups (like comparing one group that gets the program to a control group).
- If your program is short and has a defined start and finish, simpler designs like pre- and post-tests on one group are typically used.
- If the program is ongoing or has no clear start and end point for participants, you'll likely need a design that tracks each individual's progress or their performance level before the program began.

- **Interrupted Time Series (ITS) Analysis:**

  This is a statistical method used when you have long-term data collected both before and after a program or policy change is introduced. It helps see if the program caused a significant change in the trend.

3. Time constraints

- The time frame for an evaluation is often determined by several practical limitations.
- First, the evaluation should ideally align with the **program's regular cycle** or schedule to make logical sense. Second, **funding deadlines** are critical; pilot project evaluations must be finished quickly enough to prove success and secure future financing.
- Finally, the timing must also accommodate the **schedules of participants** and the **availability of the professional evaluation team**.

4. Using Evaluation Results in Health

- The final factor in selecting an evaluation design is determining what will be done with the evaluation results.
- In the health sector, data gathered can go beyond simply satisfying grant requirements.
- For example, the results can also inform future public health campaigns by highlighting which strategies (like mobile clinics versus fixed locations) are most effective for reaching vulnerable populations. Crucially, the credibility and influence of these findings—how strongly they support future decisions—are directly tied to the rigor of the research design originally selected.

Steps in Selecting an Evaluation Design

Step 1: Define Purpose and Context Checklist

The evaluation must be useful, feasible, ethical, and accurate. Define:

- **Purpose:** Why is the evaluation being conducted? (e.g., accountability, learning, improvement).
- **Context:**
  - What is being evaluated? (Program, project, policy).
  - What stage is it in? (Pilot, maturity).
  - What are the political and organizational constraints?
- **Key Questions:** What specific questions must the evaluation answer?
- **Resources:** What are the budget, timeline, and staff available?

Step 2: Analyze Evaluation Design Framework

- 3 factors that have major influence on available design options
  - Point in which the program and evaluation take place
  - The number and timing of data collection
  - Is a comparison group practical or available

Step 3: Identifying List of Potential Designs

- Based on the purpose and context, identify a few viable designs from the major categories:
- **Impact Designs:** Focus on causality (e.g., **Randomized Control Trials (RCTs)**, Quasi-Experimental Designs).
- **Process Designs:** Focus on implementation and operation (e.g., **Formative Evaluation**, Process Tracing).
- **Descriptive Designs:** Focus on "what is" (e.g., **Surveys, Case Studies**).
- **Mixed Methods Designs:** Combine quantitative and qualitative approaches.

Step 4: Considering Methodological Approach

- Determine the *type* of data and methods best suited to answer the evaluation questions:
- **Qualitative (The "Why"):** In-depth understanding, context, experiences (e.g., Interviews, Focus Groups, Observation).
- **Quantitative (The "What/How Much"):** Measurable data, statistical analysis, generalizability (e.g., Surveys, Administrative Data, Experiments).
- **Mixed Methods:** Strategic combination (e.g., using qualitative data to explain quantitative findings).

Step 5: Strengthening Basic Design

- Once a core design is selected (e.g., a simple pre/post-test), strengthen its rigor and feasibility:
- **Sampling:** Ensure the sample is representative or appropriate for the scope (e.g., stratified sampling, purposeful sampling).
- **Data Collection:** Select reliable and valid instruments (e.g., standardized scales, robust interview protocols).
- **Ethical Review:** Ensure the design respects privacy, consent, and potential risks to participants.
- **Pilot Testing:** Conduct a small-scale test of the design and instruments.

Step 6: Analyze Evaluability

- **Program Readiness:** Can the program's objectives, activities, and intended outcomes be clearly defined and measured? (If not, a design focused on program refinement may be needed first).
- **Data Availability:** Is the necessary data accessible, complete, and reliable?
- **Attribution/Contribution:** Is it possible to realistically isolate the program's effect given the timeline and external factors?

Step 7: Present the Design Option to Clients and Stakeholders

Present a concise proposal focused on how the design meets their needs:

- **The Recommended Design:** Clearly name the design and its rationale.
- **Link to Questions:** Explain exactly how this design will answer the key evaluation questions.
- **Trade-offs:** Discuss the pros and cons (e.g., RCTs offer high causal rigor but are time-consuming and expensive; Case Studies offer rich detail but low generalizability).
- **Resources:** Present the budget and timeline required for execution.

Step 8: Agreement on the Evaluation Design

- **Formal Approval:** Seek formal sign-off from the primary client/governance body.
- **Clarity on Scope:** Ensure all parties agree on the scope, limitations, data ownership, and reporting requirements.
- **Final Documentation:** Document the final agreed-upon design, methodology, and work plan, moving into the execution phase.

Reliability and Validity of Monitoring and Evaluation

Core Concepts: The Target Analogy

Reliability = Consistency

Reliability refers to the extent to which a measurement gives results that are very consistent. If you measure the same thing twice, do you get the same number?

Validity = Accuracy

Validity refers to the extent to which a test measures what it claims to measure. Are you hitting the bullseye, or just hitting the same spot off-target?

Reliability vs. Validity

Feature	Reliability (Consistency)	Validity (Accuracy)
What does it tell you?	The extent to which results can be reproduced when repeated under same conditions.	The extent to which results really measure what they are supposed to measure.
Assessment	Checking consistency across time, observers, and parts of the test itself.	Checking correspondence to established theories and other measures.
Relationship	A reliable measurement is not always valid (reproducible but incorrect).	A valid measurement is generally reliable (accurate results are reproducible).

"High reliability is one indicator that a measurement is valid. If a method is not reliable, it probably isn't valid."

Reliability and Validity

- Reliability and validity are concepts used to evaluate the quality of research.
- They indicate how well a method, technique or test measures something.
- Reliability refers to how consistently a method measures something. If the same result can be consistently achieved by using the same methods under the same circumstances, the measurement is considered reliable. 2

Reliability (consistency)

Reliability refers to the consistency of measurement that is, how consistent test scores or other evaluation results are from one measurement to others.

-Gronland and Linn

Reliability refers to the consistency of scores obtained by the same individuals when re-examined with the same test on different occasions or with different sets of equivalent items or under variable examining conditions.

-Anatasl

- **Internal consistency reliability**

➢Split half method

- **External consistency reliability**

➢Test-retest method

➢Inter-rater or inter-observer method

➢Parallel forms method

	INTERNAL	EXTERNAL
Focus	**Internal consistency** of the measurement tool.	**Stability** of results across different times/observers.
Core Question	Are all questions/indicators consistently measuring the *same* underlying concept? (e.g., Are all 10 questions truly measuring "Health Seeking Behavior"?)	Will the measurement yield the *same* result if conducted again by a different person or at a later time?
Measurement Type	**Single-Administration** (Data collected at one point in time).	**Multiple-Administration** (Data collected across time or by different people).
Statistical Test	**Cronbach's Alpha** (most common for scales/indices).	**Test-Retest Correlation** or **Inter-Rater Reliability** (Kappa statistic).
Example	**Health Knowledge Score** calculated from multiple questions	Monitoring the condition of village water taps after a water, sanitation, and hygiene (WASH) program intervention

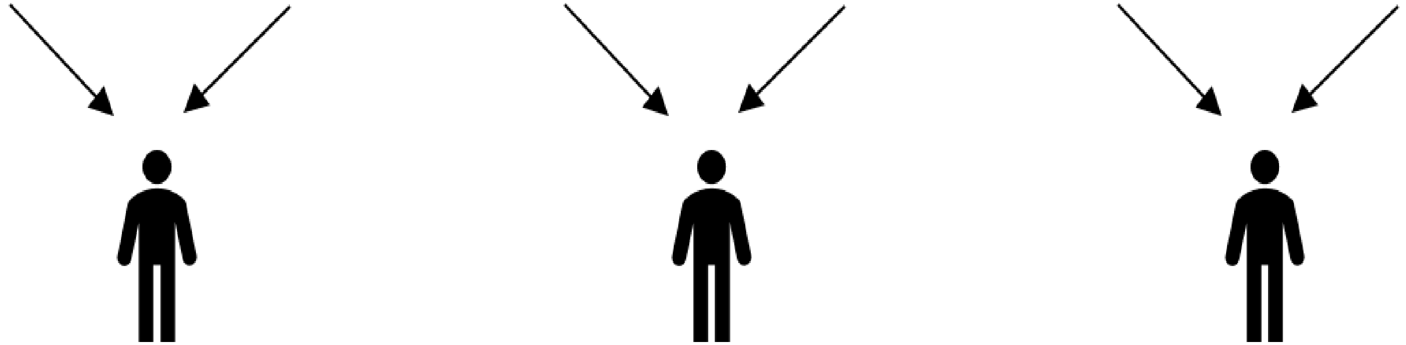
Split Half Method

- The split-half method assesses the internal consistency of a test, such as psychometric tests and questionnaires. There, it measures the extent to which all parts of the test contribute equally to what is being measured.
- This is done by comparing the results of one half of a test with the results from the other half.
- A test can be split in half in several ways, e.g. first half and second half, or by odd and even numbers. If the two halves of the test provide similar results this would suggest that the test has internal reliability.

Half 1	Half 2
Question 1	Question 2
Question 3	Question 4
Question 5	Question 6
Question 7	Question 8
...	...
...	...
Question 99	Question 100

Half 1	Half 2
Question 1	Question 2
Question 3	Question 4
Question 5	Question 6
Question 7	Question 8
...	...
...	...
Question 99	Question 100

Half 1	Half 2
Question 1	Question 2
Question 3	Question 4
Question 5	Question 6
Question 7	Question 8
...	...
...	...
Question 99	Question 100



- The split-half method is a quick and easy way to establish reliability. However, it can only be effective with large questionnaires in which all questions measure the same construct (indicator variable that measures a characteristics) This means it would not be appropriate for tests which measure different constructs.

Example: A large questionnaire assessing maternal health awareness in rural area. If odd-numbered and even-numbered questions yield similar results, the tool has internal reliability.

HIV/AIDS awareness survey among adolescents in Kathmandu Valley—splitting the questionnaire into two halves and comparing consistency.

Measuring Quality of Life in Chronic Disease Patients:

A survey section on quality of life (QoL) for Tuberculosis patients includes several items, such as "I feel hopeful about the future," and "I can perform daily tasks easily."

Measure: Cronbach's Alpha (α) is used.

High Reliability: A high alpha value (typically > 0.70) indicates that all these items reliably measure the common QoL construct.

Example : Calculate the split half coefficient of the ten-question questionnaire using a Likert scale (1 to 7) given to 15 people whose results are shown in Figure.

	A	B	C	D	E	F	G	H	I	J	K
3		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
4	1	3	2	4	1	4	5	1	4	3	2
5	2	5	4	6	3	5	1	2	5	5	5
6	3	7	3	5	5	6	4	3	3	6	5
7	4	2	4	3	3	2	2	2	6	6	3
8	5	6	5	6	5	4	4	4	6	7	5
9	6	1	2	3	4	2	5	2	3	5	2
10	7	4	3	5	3	5	2	1	2	4	3
11	8	2	4	4	3	7	2	2	4	6	5
12	9	6	5	5	5	1	6	5	2	3	2
13	10	5	4	6	4	3	5	3	1	7	2
14	11	4	3	5	2	5	3	2	3	4	1
15	12	4	2	3	2	5	3	4	2	3	4
16	13	5	5	6	2	3	3	2	4	4	5
17	14	6	5	5	4	2	2	1	4	6	5
18	15	4	3	5	5	3	1	3	2	3	4

	K	L	M	N	O	P	Q	R
22		Odd	Even					
23	1	15	14		Correlation coefficient			0.537058
24	2	23	18		Spearman-Brown correction			0.698813
25	3	27	20					
26	4	15	18		Real Statistics function			0.698813
27	5	27	25					
28	6	13	16					
29	7	19	13					
30	8	21	18					
31	9	20	20					
32	10	24	16					
33	11	20	12					
34	12	19	13					
35	13	20	19					
36	14	20	20					
37	15	18	15					

1.Pearson Correlation Formula

$r = r12$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

2. Spearman–Brown Prophecy Formula

Because each half is only half the length of the full test, we must adjust the reliability estimate.

$$r_{SB} = \frac{2r_{12}}{1 + r_{12}}$$

the most important formula in split-half reliability

Example Scenario

Topic: ***Assessing Health-Promoting Behaviors among University Students***
Tool: ***A 10-item Likert-scale questionnaire***
Construct: *Health-promoting lifestyle*

1.Prepare the Questionnaire

You create a 10-item Likert scale (1 = Strongly disagree to 5 = Strongly agree).

2. Split the Items into Two Halves

Two common ways:

Odd–Even Split

- Half A: Items 1, 3, 5, 7, 9
- Half B: Items 2, 4, 6, 8, 10

First–Second Half Split

Half A: Items 1–5 and Half B: Items 6–10

Odd–even is preferred because it balances content.

3.Compute Scores

Calculate total score for Half A

Calculate total score for Half B

4. Correlate the Two Halves

Use Pearson correlation between Half A and Half B scores.

Example result:

r=0.72

5. Apply Spearman–Brown Prophecy Formula

estimate full-scale reliability:

Reliability=2r/1+r

=2*(0.72)/{1+0.72}=0.84

Interpretation

A reliability of 0.84 indicates good internal consistency

- *Health-Promoting Lifestyle Questionnaire (10 items)*
- ***Response options:***
  1 = Strongly disagree
  2 = Disagree
  3 = Neutral
  4 = Agree
  5 = Strongly agree

How to Use This Tool for Split-Half

- Odd items (1,3,5,7,9) → Half A
- Even items (2,4,6,8,10) → Half B

Compute correlation → Apply Spearman–Brown → Report reliability.

Item No.	Questionnaire Item
1	I regularly engage in at least 30 minutes of physical activity.
2	I eat fruits and vegetables every day.
3	I avoid smoking and tobacco products.
4	I drink at least 6–8 glasses of water daily.
5	I get at least 7 hours of sleep each night.
6	I manage stress through relaxation or mindfulness techniques.
7	I maintain good personal hygiene practices.
8	I limit my intake of junk food and sugary drinks.
9	I seek medical care when I notice unusual symptoms.
10	I wash my hands before eating and after using the toilet.

Test-retest reliability

- ✓Test-retest reliability measures the consistency of results when you repeat the same test on the same sample at a different point in time.
- ✓ You use it when you are measuring something that you expect to stay constant in your sample.

Example : Measuring blood pressure of hypertensive patients in Pokhara at two different times using the same digital sphygmomanometer. If results are consistent, the tool shows good reliability.

In a study validating the **Caregiver Knowledge of Child Development Inventory (CKCDI)** in Nepal, researchers interviewed caregivers twice within a few weeks.

Goal: To ensure that a mother's knowledge score.

Why it’s important ?

- Many factors can influence your results at different points in time: for example, respondents might experience different moods, or external conditions might affect their ability to respond accurately.
- Test-retest reliability can be used to assess how well a method resists these factors over time.
- The smaller the difference between the two sets of results, the higher the test-retest reliability.

How to measure it?

- To measure test-retest reliability, you conduct the same test on the same group of people at two different points in time.
- Then you calculate the correlation between the two sets of results.

Interrater reliability

- Interrater reliability (also called interobserver reliability) measures the degree of agreement between two or more independent observers or raters when using the same instrument or procedure to measure the same event. You use it when data is collected by researchers assigning ratings, scores or categories to one or more variables.
- To measure interrater reliability, different researchers conduct the same measurement or observation on the same sample.
- Crucial for subjective measurements and observational data.

Observational Study of Handwashing Compliance:

Two M&E officers are assigned to observe healthcare workers' compliance with the "Five Moments for Hand Hygiene" protocol in a Kathmandu hospital.

Procedure: Both officers independently record their observations using the same checklist.

High Reliability: If both officers consistently record the same compliance rates for the same observed staff, the observation tool and training are reliable.

- **Supervision Performance Assessment**: Studies, such as those focusing on the Supervision Performance Assessment and Recognition (SPARS) indicators for medicines management in public health facilities, have specifically looked into IRR in the Nepali context. These studies assess how consistently different supervisors rate the same facility.
- **Routine Health Information Systems (HMIS):** Ensuring consistency in data reporting from diverse facilities (hospitals, health posts, FCHV data) across varied geographic areas (Terai, Hills, Mountains) is paramount.

Parallel forms reliability

- Parallel forms reliability measures the correlation between two different but equivalent versions of a test or survey. You use it when you have two different assessment tools or sets of questions designed to measure the same thing.
- The most common way to measure parallel forms reliability is to produce a large set of questions to evaluate the same thing, then divide these randomly into two question sets.
- Eg: evaluating the impact of a sanitation campaign on the knowledge of safe water practices among residents of a municipality in the Kathmandu Valley.

Evaluation of Health Training Modules:

A public health project develops two versions (Form A and Form B) of a post-training assessment to evaluate the learning outcomes of Female Community Health Volunteers (FCHVs) on pneumonia management.

Procedure: Half the FCHVs take Form A, the other half take Form B.

High Reliability: If both forms yield similar average scores, the forms are considered equivalent and reliable measures of the training's impact.

What is my methodology?	Which form of reliability is relevant?
Measuring a property that you expect to stay the same over time.	Test-retest
Multiple researchers making observations or ratings about the same topic.	Interrater
Using two different tests to measure the same thing.	Parallel forms
Using a multi-item test where all the items are intended to measure the same variable.	Internal consistency

Why it Matters in Nepal?

- ✓Reliable data is crucial for scarce resource allocation, justifying donor funding, and building trust in the effectiveness of health interventions across diverse geographical and cultural settings.
- ✓In the context of Nepal, investing in proper training, tool translation, and strict supervision is essential to overcome context-specific challenges.
- ✓ Reliable data→ Better Evidence→ Improved Public Health Outcomes.

Challenges in Achieving Reliability

- Human Error: Observer bias or inconsistent practices can affect reliability.
- Environmental Factors: External conditions, such as distractions during testing, can influence results.
- Time Constraints: Limited time for instrument development may compromise reliability.
- Complex Constructs: Measuring abstract concepts like emotions or attitudes can pose challenges.

Validity (accuracy)

- Truthfulness : Does the test measure what it purpose to measure?
- Validity refers to how accurately a method measures what it is intended to measure. If research has high validity, that means it produces results that correspond to real properties, characteristics, and variations in the physical or social world.
- High reliability is one indicator that a measurement is valid. If a method is not reliable, it probably isn't valid.

Internal and external validity

- Internal and external validity are concepts that reflect whether or not the results of a study are trustworthy and meaningful.
- While internal validity relates to how well a study is conducted (its structure), external validity relates to how applicable the findings are to the real world.
- Internal Validity is paramount for M&E: It confirms if the public health investment worked (causality).
- External Validity is crucial for policymakers: It determines if the program can be scaled up nationally or adapted to other districts.

- Best Practice in Nepal: Design M&E studies that are rigorous enough to establish cause-and-effect (Internal Validity) while being mindful of sampling and setting to ensure findings are relevant to the diverse reality of the country (External Validity).

Internal Validity

- Internal validity is the extent to which a study establishes a trustworthy cause-and-effect relationship between a treatment and an outcome.
- Internal validity depends largely on the procedures of a study and how rigorously it is performed.
- The less chance there is for "confounding" in a study, the higher the internal validity and the more confident we can be in the findings.
- Confounding refers to a situation in which other factors come into play that confuses the outcome of a study.

Example: A study testing the effect of a nutrition education program on reducing anemia among adolescent girls in Chitwan. Randomization and blinding improve internal validity.

Evaluating the impact of a smoking cessation program in Dharan—ensuring confounding factors like alcohol use are controlled.

External validity

- **External validity** is the measure of **generalizability**, determining whether the results found in a specific study or sample can be accurately applied to the broader population, different locations, or real-world settings.
- It answers the critical question: *"Will the success we observed in this controlled pilot also happen when we scale the program up or move it to a new region?"*
- Without high external validity, an intervention might look effective on paper (or in a test group) but fail when implemented in normal, daily conditions.

Example: Findings from a vaccination coverage survey in Lalitpur applied to other urban municipalities in Nepal.

Results of a sanitation intervention in rural Terai replicated in hill districts to test generalizability.

Factors That Improve Internal Validity

- ✓**Randomization** refers to randomly assigning participants to treatment and control groups, and ensures that there is not any systematic bias between groups.
- ✓**Random selection** of participants refers to choosing your participants at random or in a manner in which they are representative of the population that you wish to study.
- ✓**Blinding** in a study refers to participants and sometimes researchers being unaware of what intervention they are receiving (such as by using a placebo in a medication study) to avoid this knowledge biasing their perceptions and behaviors and thus the outcome of the study.

- ✓ **Experimental manipulation** refers to manipulating an independent variable in a study (for instance, giving smokers a cessation program) instead of just observing an association without conducting any intervention (examining the relationship between exercise and smoking behavior).
- ✓ **Study protocol** refers to following specific procedures for the administration of a treatment so as not to introduce any effects of, for example, doing things differently with one group of people versus another group of people.

Factors That Threaten Internal Validity

- ✓**Confounding** refers to a situation in which changes in an outcome variable can be thought to have resulted from some third variable that is related to the treatment that you administered.
- ✓**Historical events** may influence the outcome of studies that occur over a period of time. Examples of COVID-19 lockdowns affecting ongoing tuberculosis treatment adherence studies.

- ✓ **Attrition** refers to participants dropping out or leaving a study, which means that the results are based on a biased sample of only the people who did not choose to leave (and possibly who all have something in common, such as higher motivation. Mothers dropping out of a longitudinal maternal health study in remote area due to migration.
- ✓ **Experimenter bias** refers to an experimenter behaving in a different way with different groups in a study, which leads to an impact on the results of this study (and is eliminated through blinding).

✓**Maturation** refers to the impact of time as a variable in a study. If a study takes place over a period of time in which it is possible that participants naturally changed in some way (grew older, became tired), then it may be impossible to rule out whether effects seen in the study were simply due to the effect of time. Children naturally growing more resistant to disease

External Validity

- External validity refers to how well the outcome of a study can be expected to apply to other settings.
- In other words, this type of validity refers to how generalizable the findings are.
- For instance, do the findings apply to other people, settings, situations, and time periods?

Factors That Improve External Validity

- **Inclusion and exclusion criteria** should be used to ensure that you have clearly defined the population that you are studying in your research.
- **Replication** refers to conducting the study again with different samples or in different settings to see if you get the same results.
- **Field experiments** can also be used in which you conduct a study outside the laboratory in a natural setting.

Factors that threaten External Validity

- **Situational factors** such as time of day, location, noise, researcher characteristics, and how many measures are used may affect the generalizability of findings.Noise and crowding during immunization camps in rural health posts affecting responses in satisfaction surveys.
- **Selection bias** refers to the problem of differences between groups in a study that may relate to the independent variable (once again, something like motivation or willingness to take part in the study, specific demographics of individuals being more likely to take part in an online survey). Online surveys on mental health during the pandemic mostly reaching urban youth with internet access, not rural population.

References

1. 1.https://researchmethod.net/reliability/
2. https://www.scribbr.com/methodology/reliability-vs-validity/
3. https://www.scribbr.com/methodology/internal-vs-external-validity/
4. https://www.statology.org/split-half-reliability/
5. https://jnma.com.np/jnma/index.php/jnma/article/view/8952
6. He Q, Shi JP. [Realization of design regarding experimental research in the clinical real-world research]. Zhonghua Liu Xing Bing Xue Za Zhi. 2018;39(4):519-23.
7. McCauley E, Novins DK. Editorial: Research in Real-World Settings: Challenging the Limits of Experimental Design. J Am Acad Child Adolesc Psychiatry. 2020;59(6):697-8.
8. https://www.slideshare.net/slideshow/experimental-research-design-250930326/250930326

Images sources

1. https://www.scienceforsport.com/reliability/
2. https://chwcentral.org/nepals-community-health-worker-system/
3. https://clinicone.com.np/full-body-checkup-female-bkt/


Thank You...