Measurement Issues in Evaluation

Evaluation Designs

Selecting an Evaluation Design

- How much time do you have?
- What financial resources are available?
- How many participants can be included?
- Qualitative or quantitative data (or both)?
- Do you have data analysis skills?
- Do you want to generalize your findings to other populations?
Selecting an Evaluation Design

- Are stakeholders concerned with validity and reliability?
- Can you randomize participants into experimental and control groups?
- Do you have access to a comparison group?

Experimental Approaches

- Traditional type of research
- Only type of research performed in the sciences
- Future-focused
- Involves manipulation of the independent variable(s)

Types of Experimental Designs

- Non-experimental
- Quasi-experimental
- Experimental
Research Design Notation

- (R) = Random assignment
- X = Treatment
- O = Observation

Non-Experimental

- Primitive designs
- Designed only to show the effects of a treatment
- Do not control for internal validity

Non-Experimental

- No random assignment of subjects
  - Pretest-posttest design
    - Test administered, followed by the program, then the same test administered
      \[ O_1 \quad X \quad O_2 \]
  - Time series design
    - A series of observations to establish baseline, followed by the program then test administered
      \[ O_1 \quad O_2 \quad O_3 \quad X \quad O_4 \quad O_5 \quad O_6 \]
Quasi-Experimental

- Better than non-experimental
- When subjects cannot be randomly selected or randomly assigned to treatment groups
- Controls some threats to internal validity

Quasi-Experimental Designs

- Promote external validity & maintain internal validity

1. Pretest-posttest design
   - O₁ X O₂ (experimental)
   - O₃ O₄ (control)

2. Time series design
   - O₁ O₂ O₃ X O₄ O₅ O₆ (experimental)
   - O₇ O₈ O₉ O₁₀ O₁₁ O₁₂ (control)

Experimental

- Recommended
- Random selection of subjects from the population
- Random assignment of subjects to treatment groups
- Controls threats to internal and external validity
Experimental Designs

Better control of research shortcomings

- Pretest-posttest design
  - Random assignment to groups, test all groups, program to some groups then test all groups again
    
    (R) $O_1$ $X$ $O_2$ (experimental)
    (R) $O_3$ $O_4$ (control)

- Posttest only design
  - Random assignment, then program to some groups, then test all groups
    
    (R) $X_1$ $O_1$ (experimental)
    (R) $X_2$ $O_2$ (experimental)
    (R) $O_3$ (control)

3. Time series design

    (R) $O_1$ $O_2$ $O_3$ $X$ $O_4$ $O_5$ $O_6$
    (R) $O_7$ $O_8$ $O_9$ $O_{10}$ $O_{11}$ $O_{12}$

There are many different possible designs…

Figure 14: Intervention designs

- Pretest-posttest design
  - Experimental group
    
    $O_1$ $O_2$
  - Control group
    
    $O_3$ $O_4$

- Posttest only design
  - Experimental group
    
    $X$ $O_1$
  - Control group
    
    $O_2$

- Time series design
  - Experimental group
    
    $O_1$ $O_2$ $O_3$ $X$ $O_4$
  - Control group
    
    $O_5$ $O_6$ $O_7$ $O_8$ $O_9$ $O_{10}$ $O_{11}$ $O_{12}$

Note: 
- $O$ = random assignment
- $X$ = intervention
- $+$ = posttest
Internal Validity

The degree to which the program caused the change that was measured? In other words, were the changes in the participants due to your program or something else?

Threats to Internal Validity

- History
- Maturation
- Testing
- Instrumentation
- Statistical Regression
- Selection
- Experimental Mortality
- Diffusion or imitation of treatments
- Compensatory equalization of treatments
- Compensatory rivalry
Threats to Internal Validity

- **History:**
  - Something happens during the program to influence the results (e.g. During your program on healthy eating, the low carb phenomenon hits)

- **Maturation:**
  - People may change because they grow older, wiser, stronger

- **Testing:**
  - Pre-test/Post-test results are different because people are familiar with the tests used

- **Instrumentation:**
  - Change in measurement procedures (e.g. M-F measures of caloric intake vs. M-Su)

- **Statistical Regression:**
  - Extremely high or low scores on the pre-test will come back to the center on the post-test

- **Selection:**
  - Differences between the experimental and control groups due to lack of randomization

- **Mortality/Participant Attrition:**
  - Participant drop-out. Who is left standing at the end? Has the make-up of the group changed significantly?

- **Diffusion or Imitation of Treatments:**
  - Control group interacts with experimental group and imitates their behavior

- **Compensatory Equalization:**
  - It’s not fair!! Control group whines about not getting the treatment or program
Threats to Internal Validity

- Compensatory Rivalry:
  - When the control group is seen as the “underdog” and tries harder on the post-test (a.k.a. Avis Effect)
- Resentful Demoralization (more whining):
  - One treatment group may perceive another group to be getting a better treatment or program so they don’t try as hard or give up

Methods of Control

Randomization (controls many threats to internal validity)
- Random selection of participants from the target population
- Random assignment to treatment or control groups (experimental)
- Random assignment of types of treatment or no treatment to groups (quasi-experimental)

Measurement Validity
Measurement Validity

- Validity
  - The degree to which an instrument measures what the evaluator wants it to measure
  - Types:
    - Face
    - Content
    - Criterion
    - Construct

- Face validity
  - Extent to which the instrument appears to be measuring what it is supposed to measure

- Content validity
  - Extent to which an instrument samples items from the universe of content desired

- Criterion validity
  - Extent to which an instrument correlates with another more accurate instrument (the criterion)

- Construct validity
  - Extent to which the measure correlates with other measures in predicted ways, but for which no true criterion exists
Reliability

- An empirical estimate of the extent to which an instrument produces the same result, applied once or two or more times

Biases to Valid and Reliable Measurement

- Guinea pig effect
  - People who are aware that they are being measured may respond in uncharacteristic ways

- Role selection
  - Awareness of being measured may influence people to feel that they have to play a special role

Biases to Valid and Reliable Measurement

- Measurement as a change agent
  - The act of taking a measurement may affect the subsequent behavior or those being measured.

- Response sets
  - People respond to questionnaires and interview in a socially desirable, predictable way that have little or nothing to do with the questions posed.
Biases to Valid and Reliable Measurement

- Interviewer effects
  - Characteristics of the interviewer may affect respondents' receptivity and answers.
- Changes in the instrument
  - If an instrument is used more than once, a learning effect is possible.
  - Interviewers may become more proficient at or tired or implementing an interview schedule the longer they use it.

Biases to Valid and Reliable Measurement

- Population restrictions
  - The method of data collection may impose restrictions on the population to which the results can be generalized.
- Population stability over time
  - An instrument administered at different times may not collect the same data on different populations

Biases to Valid and Reliable Measurement

- Population stability over areas
  - The same way of collecting data in two different geographic areas may assess different types of people
- Content restrictions
  - A limited range of data can be reported by each method
Biases to Valid and Reliable Measurement

- Stability of content over time
  - If a program restricts a study to naturally occurring behavior, the content of the studied phenomenon may vary over time.
- Stability of content over an area
  - A program may be uniform in content throughout the area in which it is applied.

Data Collection Methods

Types of Variables

- Demographic
- Cognitive
- Skill
- Psychosocial
- Behavioral
- Environmental
- Health
- Cost
Obtrusive Measures

- Self-completion questionnaires
- Self-completion main surveys
- Self-completion diaries and logs
- Face-to-face interviewing
- Telephone interviewing
- Direct observation

Unobtrusive Measures

- Abstraction of existing medical and clinical records
- Clinical or physiological measures

External Validity
External Validity

- Can the findings be generalized beyond the subjects or settings studied in the experiment?
- If we do this again with some other group, will it work again?

Threats to External Validity

- Social desirability
  - Subject responds the way he/she thinks the experimenter wants them to
- Expectancy effect
  - When people act as you expect them to act (Rosenthal effect)

Threats to External Validity

- Hawthorne effect
  - Behavior change due to people feeling special
- Placebo effect
  - Change in behavior due to belief in the effectiveness of the treatment
Controlling Threats to External Validity

- Blind
  * Participants
- Double blind
  * Participants and investigators
- Triple blind
  * Participants, investigators, evaluators