

Measurement

Measurement assumptions in common sense statements

'I think attractive people are more successful because they're more likely to be selected at interviews and to be given more attention generally'

'No. It could be that more attractive people develop better social confidence earlier on in life and that's what gets them through interviews and the like'

Research of such statements requires operationalization of concepts like

'attractiveness'
'success'
'self-confidence'

In some way values must be attached to different levels of attractiveness etc.

Quantitative and qualitative differences

'Helen is more artistic than Claire'

'George is a contemplative type whereas Rick is practical, energetic and impulsive'

How do we **know** George is contemplative, while Rick is practical, energetic and impulsive' ?

In order to make the judgment we need to compare some things they do (how strongly and how often) with their occurrence in others

In order to demonstrate a difference we need some numerical measure

SCALE TYPES

Nominal level of measurement

Mutually exclusive categories
qualitative

Ordinal level of measurement

= Nominal scale + Ordered categories
quantitative, ranked

Interval level of measurement

= Ordinal scale + Equal spacing of categories
quantitative, ranked, equal intervals between categories

Ratio level of measurement

= Interval scale + Absolute zero
quantitative, ranked, equal intervals between categories with a true zero point

Nominal scale

Male - Female
Red, Green and Blue objects
Roman noses - Other noses
Attractive - Average - Unattractive

Each item is compared with some learnt concept

Difficulties may arise in categorizing a person in for example

Smoker - Non-Smoker
Introvert - Extrovert
Optimist - Pessimist

Items are bunched together on the basis of some common feature

Nominal scales examples

Categorization table

Category	1	2	3	4	5
	Students	Teaching staff	Non-Teaching staff	Visitors	Other
Frequency	650	34	43	17	2

Number of voters by political party

Communist	Conservative	Labor	SDP	SLP	Other
243	14678	15671	2356	4371	567

Number of people smoking an average of N cigarettes per day

N =	None	1-5	6-10	11-20	21-30	31-40	41+
	65	45	78	32	11	4	3

Ordinal scales

Ordinal numbers represent position in a group

Who came 1st, 2nd, 3rd etc. but not how far apart

General knowledge scores

Person	Score	Rank of score
1	18	5.5
2	25	7
3	14	1
4	18	5.5
5	15	3
6	15	3
7	15	3
8	29	8

'Gray horses are slower than brown horses'

Nominal data

	Gray	Brown
Finished in top 10	3	7
Finished in last 10	7	3

Ordinal data

Color of horse	Gray	Brown
	1	4
	2	5
	3	6
		7
	11	8
	12	9
	13	10
	14	
	15	18
	16	19
	17	20

Interval sales

Intervals have equal sizes

10 to 15 minutes is the same interval as 20 to 25 minutes
 The interval 30° to 40° is twice the interval -10° to -5°

Philosophical discussion: Is IQ a true interval measure?

Example IQ

Child	IQ
Jane	100
Jacky	80
Joanna	120

Jane is as far ahead of Jacky as Joanna is ahead of Jane
 Are two scores of, say, 110 equal?

Danger of REIFICATION: it is not because we apply numbers to IQ's that intelligence exists as something with quantity

Reduction of data from Interval to Ordinal level

Reaction time	Rank
0.067	1
0.078	3
0.091	5
0.089	4
0.076	2

Intimacy rating	Rank
7	4
6	2.5
5	1
6	2.5
9	5

Reduction of data from Interval to Nominal level

# anxiety indicators	High	Low
competitive children	14	10
	21	6
	7	13
	13	5
	18	11

mean= 11.8

	Level of competitiveness	
	High	Low
Anxiety > mean	4	1
Anxiety < mean	1	4

Ratio scales

True zero

Time, Distance, Temperature in °Kelvin, Most measures of physical qualities

Time taken over course (sec)	Grey	Brown
	123	132
	124	136
	125	136
		137
	143	139
	143	142
	143	142
	144	
	146	153
	146	154
	147	156

Scaling

A construct is generally not directly observable / measurable

Researcher must assign scores to people or objects to use as measures for a construct by referring to observable attributes that are somehow related to the construct

Scale = a set of categories or range of scores on a variable

Scaling = process of assigning scores to objects to yield a measure of a construct

Scaling techniques can be based on Judgements
 Multiple responses

Judgments

Individual in question or observers assign scores to reflect the underlying construct

- "Do you consider yourself to be very liberal, liberal, middle of the road, conservative, or very conservative"
- Raters may rate news stories as "favorable, neutral, or unfavorable"

Multiple responses

Several measurements (responses to questions) are combined into a single scale score.

- In order to assess overall 'liberalism' or 'conservatism' in a congressional representative his or her position on a number of roll-call votes may be combined to give a single score.
- Questions asked to assess a white person's attitude towards racial integration:
 "I would prefer to have blacks as well as whites in school classes"
 "Property values decline when black people move into a neighborhood"
 etc.

Advantages of multiple item scaling

- Scaling can **reduce the complexity** of the data
One combined score for "liberalism" >< individual roll-call votes
- Allows to investigate the **dimensionality** of a construct
Unidimensional requires **high correlations between variables**
eg. *votes on social welfare issues* >< *votes on human rights issues*
- Improved **reliability and validity** of measurement
Systematic error in some variables and random error are expected to cancel out

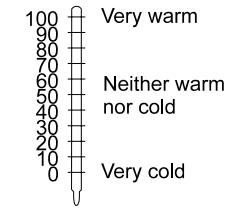
Rating scales for quantifying judgments

Judgment task

- 1) **forming a subjective impression of stimulus along the desired dimension**
(judge may not be able to consider only the desired dimension)
- 2) **translating judgment into an overt rating**
(judge's frame of reference, own attitude, context effects)

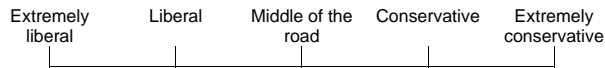
Ratings given to single object cannot be understood without knowing something about the range of objects with which the judge is implicitly comparing it

Graphic rating scales



Caution with extreme categories
Sensitive to frame of reference problem

Itemized rating scales



More clearly defined categories improve reliability

Comparative rating scales

"As compared with the total group of graduate students you have known"
Is the applicant more capable than
◊ 10 percent of them?
◊ 20 percent?
◊ 30 percent?

"Indicate to whose leadership skills that of the candidate resembles most."
Person A Person B Person C

Self-ratings versus Ratings by Others

- Self ratings are often superior but only if
- Individuals receive very clear instructions about the attribute to be rated
 - Opportunity and incentives are given to recall past behaviors
 - Individuals are motivated to give accurate ratings

Construction and use of Rating scales

HALO bias = tendency for overall positive or negative evaluations

Generosity error = overestimation of desirable qualities of people the rater likes

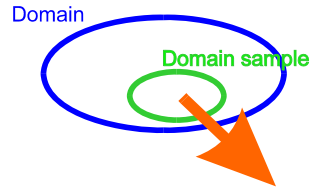
Avoidance of extreme categories

Contrast error = tendency for raters to see other people as opposite to them on a trait

Minimizing the impact of errors by training, motivating of raters, multiple raters, ..., development of the scales

MULTIPLE-ITEM SCALES

Domain = hypothetical population of all items relevant to the construct
Domain sample = (nonprobability) sample of items



More items results in a more reliable scale
Amount of variation in item content determines the reliability of the measure

Item construction

- Items must be empirically related to the construct under study
- Items must differentiate among people who are at different points along the dimension to be measured
- Avoid double-barreled items or otherwise ambiguous or vague items
- Include both positive and negatively worded items

Likert scales

Basis: probability of agreeing with favorable items increases directly with the degree of favorability of the subject's attitude

Respondents indicate a degree of agreement
Scale score = sum of itemscores (not average)

Construction

- Items that are clearly favorable or unfavorable are assembled based on a theoretical conception of the concept to be measured
- Items are administered to pilot subjects
- Subjects' scale scores are computed
- Responses are analyzed to determine which items contribute most to the reliability and validity of the measure

Likert scales

Item analysis

- correlation of individual items with scale score
- factor analysis

Use of subtle items

- Based on pure empirical relation with the studied construct
- Can partly disguise the researcher's purpose
- Difficult to find, requires huge item set and pilot sample to start with
- Evidence is found that subtle items don't help to measure more accurately

Dimensionality

Unidimensional construct: All items should correlate moderately to highly

Multidimensional construct: Scale should divide in subscales measuring different dimensions

Likert scales

Advantages

- Simpler to construct than Thurstone scale
- Can be used in many cases where Thurstone or Guttman scales cannot (e.g. multidimensional domains)
- Generally more reliable than Thurstone scale of the same length
- Range of responses makes subjects feel more comfortable

Disadvantages

- No information on subject's latitude of acceptance to measure the degree of issue involvement (>< Thurstone)
- Scale score carries no information about the exact pattern of responses (>< Guttman)
- If a scale in multidimensional then the same Likert score may derive from conceptually different contents

Construction of Thurstone scales

- Judges classify items into (± 11) categories ranking different favorabilities towards the attitude
- Scale value for each item is calculated as the average category placement by the judges
- Items for which the judges disagree are discarded as ambiguous or irrelevant
- Items representing a wide range of scale values are selected
- Selected items presented to subjects in random order
- Subjects check each statement with which they agree
- Subjects' attitude is calculated as the mean of scale values of the checked items

Thurstone scales

Advantages

- Responses offer a check on the scale's assumptions: subjects should agree only with a narrow range of items around their own position

Latitude of acceptance

The range of items a subject agrees with is related to the involvement in the issue. People who are more involved tend to agree with a narrower range of positions

Disadvantages

- Construction procedure is very heavy
- Attitude of judges interferes with the assignment of scale values
- Reliability tends to be lower than for Likert scales

Guttman scales

Items are ranked according to "difficulty" e.g. $A > B > C > D$

A subject who "passes" problem B is expected to "pass" C and D as well

The level of a subject is thought of as a point on the scale A-B-C-D

Scale score = number of items "passed" or "agreed" with

Examples:

- Economic liberalism
- Favorability to civil rights of Supreme Court justices
- Conditions in which most physicians would recommend abortion

Guttman scales

Advantages

- A single number carries complete information about the exact pattern of responses to every item (in case there is no random error)
- Provides a test for unidimensionality of the attitude (problem = error)
- Measures for reproducibility / scalability

Disadvantages

- Sensibility to error
- Unidimensionality is seen as a property of the items, whereas it is the pattern of an attitude in a specific population that should be investigated
- Unidimensional domains are rare

Semantic differential

For each concept the subject is asked to make a series of ratings on a multiple-point response scale

Example:

	Me as I am							
Fair	1	2	3	4	5	6	7	Unfair
Clean	1	2	3	4	5	6	7	Dirty
Light	1	2	3	4	5	6	7	Heavy
Large	1	2	3	4	5	6	7	Small
Passive	1	2	3	4	5	6	7	Active
Strong	1	2	3	4	5	6	7	Weak
Slow	1	2	3	4	5	6	7	Fast
Bad	1	2	3	4	5	6	7	Good

Semantic differential

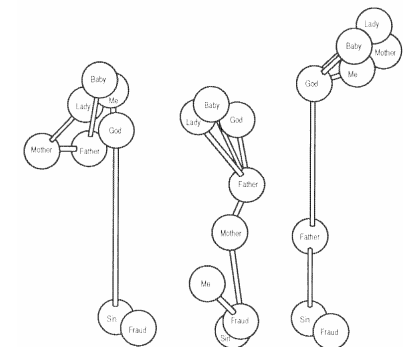
Underlying attitude dimensions

Individual's evaluation of the concept	Fair-Unfair Clean-Dirty Good-Bad Valuable-Worthless
Perception of potency or power of the concept	Large-Small Strong-Weak
Perception of activity of the concept	Active-Passive Fast-Slow Hot-Cold

Sum of scores on each dimension indicates position of the individual's attitude towards the concept

Semantic differentials have been used to measure similarity / difference between an individual's concepts of different objects (e.g. "Me as I am" <-> "Me as I would like to be")

Conceptual structure - Semantic space



Randomized experiments

Goal is to assess causality

Independent variable \Rightarrow Dependent variable

Experiments focus on independent variables that can be manipulated

Experimenter creates the levels of the experimental variables

Subject variables or organismic variables are properties of subjects that are controlled by

- holding subject variables constant (including only certain types of subjects)
- random assignment of subjects to conditions

Random assignment \rightarrow Equal groups

Random sampling \rightarrow Representative sample

Between-subjects design

Different subjects are assigned to different conditions of the independent variable

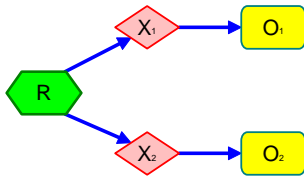
Within-subjects designs

The same subjects receive all levels of the independent variable in random order

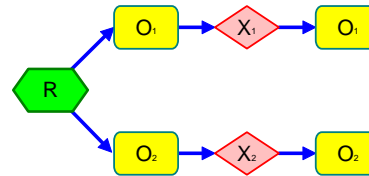
Threats to internal validity

Selection	Uncontrolled differences between test groups
Maturation	Natural processes that change subject's performance
History	Event that coincides with the treatment
Instrumental	Changes in measure procedures
Mortality	Dropout
Selection by maturation	Differences between subjects that make them change differently

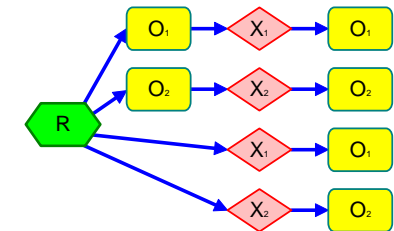
Randomized Two-Group Design



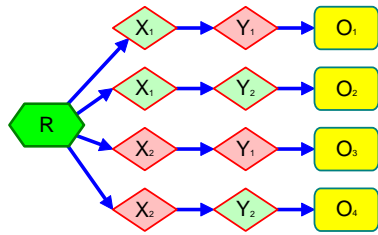
Before-After Two Group Design



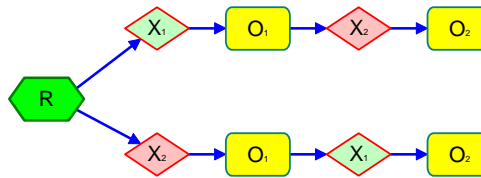
Solomon Four-Group Design



Factorial Design



Repeated Measures



Strengths and Weaknesses of Randomized Experiments

Experimental Artifacts	Subjects may behave in a way to please the researcher
External validity	Internal validity may be maximized at the expense of external validity (laboratory situation, choice of subjects)

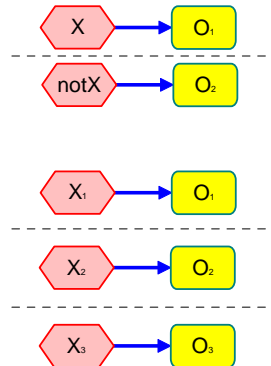
Quasi-Experimental Designs
 Survey Research Designs
 Correlational Studies

Non-Random Assignment of subjects to conditions

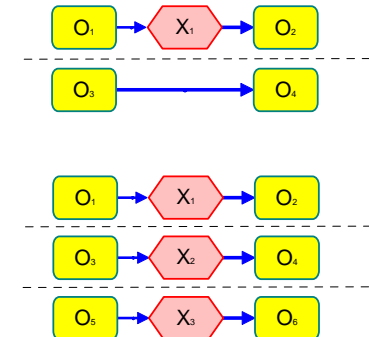
Survey research

- Sociologists collect data on representative sample of male members of the US labor force to study their training and occupational attainments
- Public opinion polling organizations conduct studies of the popularity of various presidential candidates among potential voters
- Market research organizations conduct studies of consumers to find out what kinds of soft drinks they would prefer
- Medical researchers survey the nation's population to determine the incidence of disease related characteristics
- Political scientists interview members of the US House of Representatives to monitor their reactions to increasing public attention to the ethics of public figures
- A national women's magazine asks its readers to answer a questionnaire that solicits information about occupational aspirations

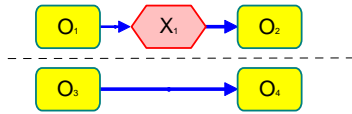
Static-Group Comparison Design



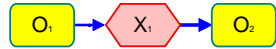
Pretest-Posttest Nonequivalent Control Group Design



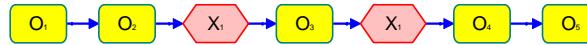
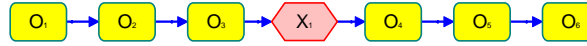
Regression-Discontinuity Design



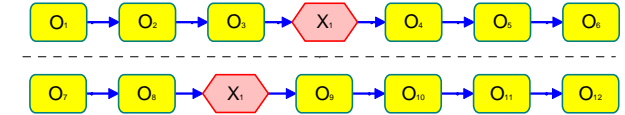
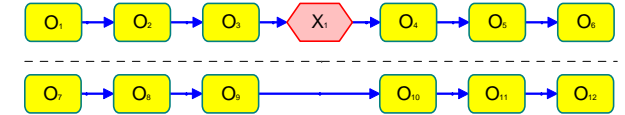
One-Group Pretest-Posttest Design



Interrupted Time-Series Design



Replicated Interrupted Time-Series Design



Sampling

External validity ?

Particularistic research → generalize research results from sample to target population

Universalistic research → test hypothesis based on a theory, applicability of theory is generalized rather than research result in sample

BASIC DEFINITIONS

Population	= aggregate set of all case that conform some set of specifications
Subpopulation	= population stratum = stratum
Population element	= single member of population
Census	= count of elements in population, determination of characteristics of the population based on info on all members
Sample	= selection of some elements of population

Representative sampling plan carries the insurance that say, 90% of the time (**confidence level**) the population estimates based on the sample differ no more than 5% (**margin of error**) from the real value

Nonprobability sampling

- Accidental Samples** select the first population elements you encounter. **danger:** underrepresentation of minorities, females, etc.
- Quota Sampling** accidental sample while taking care to have all strata represented in the sample as in the population. **danger:** "own-friends"-bias
- Purposive Samples** pick cases that are judged to be "typical" of the target population. **danger:** judgement ...

Probability sampling

- Simple Random Samples** selection based on random numbers such that each population element has equal and independent probability of being sampled
- Stratified Random Sample**
- Cluster Sample**

Mean scores of samples

Population

CASE	A	B	C	D	E	F	G	H	I	J	Mean
SCORE	0	1	2	3	4	5	6	7	8	9	4.5

Sample Means	Samples of 2 Cases	Samples of 4 Cases	Samples of 6 Cases
.5	1		
1.0	1		
1.5 - 1.75	2	2	
2.0 - 2.67	5	10	2
2.75 - 3.25	3	25	10
3.33 - 4.00	8	43	52
4.17 - 4.83	5	50	82
5.00 - 5.67	8	43	52
5.75 - 6.25	3	25	10
6.33 - 7.0	5	10	2
7.25 - 7.5	2	2	
8.0	1		
8.5	1		
No. of samples	45	210	210
Mean of s. means	4.5	4.5	4.5
% sample means > 4.00 and < 5.00	11	24	39
% sample means > 2.67 and < 6.33	60	89	98

Laboratory research

Universalistic research	Particularistic research
Basic research	Applied research
What can happen ?	What does happen ?
Manipulable independent variables	Nonmanipulable independent variables
Short timeframe	Long time frame
Awareness	Unawareness