Research Instruments, Validity dan Reliability

Trisasi Lestari - 2016
Instrument is the generic term that researchers use for a measurement device (survey, test, questionnaire, etc.).
What is measurement?
Why Measure?

- To see the differences
- To estimate the degree of relationship.
Objective assessment

Prevalence of overweight (%)

- <20
- 20–39.9
- 40–59.9
- ≥60
- Data not available
- Not applicable

* BMI ≥25kg/m²
Level of measurement

Nominal
- Numbers are assigned to objects or categories, without having numerical meaning.

Ordinal
- Objects of a set are rank ordered on an operationally defined attribute.

Interval
- Represent equal distances among attributes.

Ratio
- Ratio scales have an absolute zero point.
Subjective assessment
How to measure?
How to measure TRUST
Operationalization

- defining a concept in terms of criteria that specify how to observe, describe and measure the concept.

- the process whereby concepts can be applied on an empirical level so that they can be transformed into variables or indicators

  - Variables: symbols to which we can assign numerals or values based on measurements of the concept’s properties.

  - Indicators: indirect measures of concepts
Measurement Criteria

- Validity
- Reliability
- Generalizability
Designing Research Instruments
instrument construction
An *instrument* is a mechanism for *measuring* phenomena, which is used to *gather* and *record* information for assessment, decision making, and ultimately understanding.
FIGURE 1.1: CATEGORIES OF SOCIAL SCIENCE INSTRUMENTS.

- Behavioral rating scales
- Checklists and inventories
- Survey questionnaires
- Psychometric instruments
- Rating scales
  - Modes
    - Self-report
    - Independent rater
- Tests (teacher-made and standardized)
Test:
- a collection of items developed to measure some human educational or psychological attribute
- a *correct* answer or level of performance is anticipated

Behavioral Rating Scale:
- designed to measure an individual’s ability to complete a task or perform an activity.

Checklist
- to determine the presence or absence of an attribute and to count the prevalence of an item or event.
Inventory
- a list of objects, goods, or attributes.

Psychometric instruments
- Instruments designed to assess cognitive, affective, and physical functioning and personality traits.

Questionnaire
- designed to obtain factual information and to assess beliefs, opinions, and attitudes.
<table>
<thead>
<tr>
<th>Mode of administration</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Researcher-completed</strong></td>
<td><strong>Subject-completed</strong></td>
</tr>
<tr>
<td>- Rating scale</td>
<td>- Questionnaires</td>
</tr>
<tr>
<td>- Interview guide</td>
<td>- Self-checklist</td>
</tr>
<tr>
<td>- Tally sheet</td>
<td>- Attitude scales</td>
</tr>
<tr>
<td>- Flowchart</td>
<td>- Personality inventories</td>
</tr>
<tr>
<td>- Performance checklist</td>
<td>- Achievement test</td>
</tr>
<tr>
<td>- Time and motion logs</td>
<td>- Projective devices</td>
</tr>
<tr>
<td>- Observation forms</td>
<td>- Sociometric devices</td>
</tr>
</tbody>
</table>
Components of an Instrument

- Title
- Introduction
- Directions or Instruction
- Questions
- Selection Items / Response set
- Supply items
- Demographics
- Closing
### Continuing Education Workshop Evaluation Form

<table>
<thead>
<tr>
<th>Instructor #</th>
<th>Schedule #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Course Title:**

**Course Number:**

**Schedule Number:**

**Name of Instructor:**

Please assess each of the following statements based on the key by filling in the bubble in the column that best represents your opinion.

<table>
<thead>
<tr>
<th>SA - Strongly Agree</th>
<th>A - Agree</th>
<th>N - No Opinion or Neutral</th>
<th>D - Disagree</th>
<th>SD - Strongly Disagree</th>
<th>NA - Not Applicable</th>
</tr>
</thead>
</table>

1. The objectives for the course were clearly stated.
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

2. The instructor effectively taught the stated objectives.
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

3. The instructor used a variety of teaching strategies (two or more of the following: lecture, discussion, small group activity, visual/audio aids, individual assistance).
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

4. The workshop demonstrated how to apply the strategy or process presented.
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

5. The instructor demonstrated openness and receptivity to student needs and opinions.
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

6. Facilities and equipment (e.g., audiovisual equipment) were adequate. If not, comment below.
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

7. The subject matter was relevant to my professional needs.
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

8. What I liked best about this course was:
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

9. To strengthen the course I would suggest:
   - [ ] SA
   - [ ] A
   - [ ] N
   - [ ] D
   - [ ] SD
   - [ ] NA

10. Additional workshops, courses, or programs I would like the University Continuing Education to offer:
    - [ ] SA
    - [ ] A
    - [ ] N
    - [ ] D
    - [ ] SD
    - [ ] NA
Selecting an Instrument

Consider:

- The purpose of the study
- The research design
- Object of measurement
- Data Collection Methodology
- Resources
- Characteristics of population (potential response rate)
- Access to subjects
Questionnaire

- A self-contained and a self-administered instrument for asking questions.
- Lack the personal touch
- Extremely efficient
- Most popular
- Good questionnaire → ‘stands on its own’
Risks

Low response rates

Bias
  • Respondent bias, half-selection

Respondent honesty
  • over-report good things, and under-report bad things

Wording
  • ‘end pregnancy’ vs ‘abortion; ‘poor’ vs ‘welfare’
FIGURE 1.2: STEPS IN THE INSTRUMENT CONSTRUCTION PROCESS.

Identify the purpose and focus of the study (questions you want answered).

Obtain feedback from stakeholders to clarify the purpose and focus.

Identify the research methodology and type of instrument to use for data collection and measurement.

Begin to formulate questions or items.

Pretest items and preliminary draft (with content experts, stakeholders, potential respondents or raters).

Revise instrument based on feedback; prepare for pilot testing.

Pilot test and revise prior to final administration.

Administer instrument; analyze and report results.
Question Rules and bad examples

Clear in meaning and free of ambiguity

- “Do you do sport regularly?”
- “What is your total wealth?”

Use common everyday language, avoid jargons, abbreviations, or acronyms

- MDGs, Strategic Plan,

Use neutral language, avoid emotional, leading language

- “What do you find offensive about flag burning?”
- “Why do you think hitting children is wrong?”
Simple and easy

- “How do you rate police response time to emergency and non-emergency calls?”
- “How many cigarettes you smokes in a year?”

Asks yourself

- Does the questions answers my research question?
- Is related questionnaire existed?
- Do I need open-ended or close-ended questions?
Sample of standard questionnaires

- Generic instruments
  - COOP/WONCA charts: measure six core aspects of functional status: physical fitness, feelings, daily activities, social activities, change in health and overall health.
  - Sickness Impact Profile (SIP)/Functional Limitations Profile (FLP)

- RAND SF 36
- Duke Health Profile (DUKE)
- EuroQol
- MOS 20
- Nottingham Health Profile
- RAND General Health Perception Questionnaire (GHPQ)
Dimension specific instruments

- Barthel Index
- Index of Independence in Activities of Daily Living
- Frenchay Activities Index
- General Health Questionnaire (GHQ)
- RAND Mental Health Inventory (MHI)
- McGill Pain Questionnaire (MPQ)
Disease/condition specific instruments
- State-Trait Anxiety Inventory (STAI)
- Center for Epidemiologic Studies Depression Scale (CES-D)
- Arthritis Impact Measurement Scale (AIMS)
- Living with Asthma (AQ)
- Chronic Respiratory Disease Questionnaire (CRDQ)
- Asthma Quality of Life Questionnaire (AQLQ)
- Diabetes Health Profile IDDM (DHP 1) and NIDDM (DHP2)
- Diabetes Quality-of-Life measure (DQOL)
- EORTC Quality of Life Questionnaire
Techniques to create content of questionnaire

- Literature review
- Use available questionnaire
- Brainstorming
  - Nominal Group Technique
    - Group 5-6 people
    - Facilitator explain the purpose and problems
    - All participants write and share ideas
    - Other participant may ask for clarification
    - Repeat the brainstorming process until all ideas are collected.
    - All participant review all ideas
    - Develop a priority ranking
Creating content of questionnaire

- Snowballing / Pyramiding
  - 2 → 2+2 → 4+4 → dst

- Delphi technique
  - Researcher create the first draft
  - Collect input from experts through email/letter.
  - Experts give comments independently.
Creating content of questionnaire

- Questions Pool and Q-sort
- 60-90 questions
- Print a question in a card
- Shuffle the card
- Assess each question with a priority ranking:
  - most definitely include this item,
  - include this item,
  - possibly include this item, and
  - definitely do not include this item.
Membuat isi kuesioner

Concept Mapping

Preparation.

Generation.

brainstorming, nominal group technique, to generate statements describing activities related to the project.

Structuring.

sort the statements: Q-sort or other ranking process.

Representation.

create visual maps that reflect the relationship between the sorted items.

Interpretation.

Utilization.
### TABLE 5.3: TABLE OF SPECIFICATIONS.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Operationalized</th>
<th>Suggested Item Stems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>A. Therapy sessions are scheduled in a timely manner.</td>
<td>A. After you contacted the clinic, how long did you have to wait until you were able to meet with a therapist?</td>
</tr>
<tr>
<td></td>
<td>B. Able to meet with a therapist on short notice.</td>
<td>B.1. If a crisis arose, were you able to meet with your therapist on short notice?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B.2. If your therapist was unavailable, were you able to meet with other staff, such as a crisis worker?</td>
</tr>
<tr>
<td>Informed</td>
<td>Client is given information about diagnosis, results of evaluation, and recommendations for treatment.</td>
<td>A. To what extent were you informed about the results of testing?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. If you had questions, did your therapist explain things to you using terms that you understood?</td>
</tr>
<tr>
<td>Involved</td>
<td>Client is provided opportunity to participate in decision making about treatment.</td>
<td>When you began treatment, did your therapist work with you to develop a treatment plan that addresses your needs?</td>
</tr>
</tbody>
</table>
### Likert Scale

**Rensis Likert**  
1903 – 1981

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
<td>Very Frequently</td>
</tr>
<tr>
<td>Agree</td>
<td>Frequently</td>
</tr>
<tr>
<td>Undecided</td>
<td>Occasionally</td>
</tr>
<tr>
<td>Disagree</td>
<td>Rarely</td>
</tr>
<tr>
<td>Strongly Disagree</td>
<td>Never</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Importance</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Important</td>
<td>Almost Always True</td>
</tr>
<tr>
<td>Important</td>
<td>Usually True</td>
</tr>
<tr>
<td>Moderately Important</td>
<td>Occasionally True</td>
</tr>
<tr>
<td>Of Little Importance</td>
<td>Usually Not True</td>
</tr>
<tr>
<td>Unimportant</td>
<td>Almost Never True</td>
</tr>
</tbody>
</table>
Likert Scale Analysis

- Likert Scale: is the sum of responses on several Likert items
- Ordinal or Interval
- Descriptive
  - Median, Mode, Percentiles/quartiles, Display Distribution (bar chart)
- Non-parametric test
  - Chi-squared, Mann Whitney test, Wilcoxon signed-rank test, Kruskal-Wallis test
- Modified binomial Likert Scale
  - Chi-squared, Cochran-Q, McNemar test
Observation Checklist

Surgical Safety Checklist

**Before induction of anaesthesia**
(with at least nurse and anaesthetist)

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>Not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the patient confirmed his/her identity, site, procedure, and consent?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has the site marked?</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the anaesthesia machine and medication check complete?</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the pulse oximeter on the patient and functioning?</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the patient have a:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Known allergy?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult airway or aspiration risk?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk of &gt;500ml blood loss (7ml/kg in children)?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the patient have a:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Known allergy?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult airway or aspiration risk?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk of &gt;500ml blood loss (7ml/kg in children)?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Before skin incision**
(with nurse, anaesthetist and surgeon)

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>Not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirm all team members have introduced themselves by name and role.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirm the patient’s name, procedure, and where the incision will be made.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has antibiotic prophylaxis been given within the last 60 minutes?</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anticipated Critical Events</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To Surgeon:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What are the critical or non-routine steps?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How long will the case take?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the anticipated blood loss?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To Anaesthetist:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are there any patient-specific concerns?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To Nursing Team:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has sterility (including indicator results) been confirmed?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is essential imaging displayed?</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Before patient leaves operating room**
(with nurse, anaesthetist and surgeon)

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>Not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurse Verbally Confirms:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The name of the procedure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completion of instrument, sponge and needle counts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specimen labelling (read specimen labels aloud, including patient name)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whether there are any equipment problems to be addressed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To Surgeon, Anaesthetist and Nurse:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What are the key concerns for recovery and management of this patient?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This checklist is not intended to be comprehensive. Additions and modifications to fit local practice are encouraged.

Revised 1 / 2009 © WHO, 2009
**Pretesting**

- **Initial Pretesting**
  - Individual Interviews and Focus Groups
  - Review by Content Area Experts
  - Continue to Obtain Feedback and Revise the Project If Necessary

- **Pretesting during development**
  - Read and Reread the Items and Read the Items Aloud
  - Review by Content Area Experts
  - Review by Instrument Construction Experts
  - Review by Individuals with Expertise in Writing
  - Review by Potential Users
Questions for experts

- Was each set of directions clear (that is, the general directions at the beginning of the questionnaire and any subsequent directions provided in the body of the instrument)?
- Were there any spelling or grammatical problems? Were any items difficult to read due to sentence length, choice of words, or special terminology?
- How did the reviewer interpret each item? What did each question mean to them?
- Did the reviewer experience problems with the item format(s), or does the reviewer have suggestions for alternative formats?
- Were the response alternatives appropriate to each item?
What problems did the reviewer encounter as a result of the organization of the instrument, such as how items flowed?

On average, how long did it take to complete? What was the longest time and what was the shortest time it took to complete the instrument?

For Web-based instruments, did the respondent encounter any problems accessing the instrument from a computer or navigating the instrument once it was accessed?

Did any of the reviewers express concern about the length of the instrument, or did they report problems with fatigue due to the time it took to complete?

What was the reviewer’s overall reaction to the questionnaire?

Did they have any concerns about confidentiality or how the questionnaire would be used?

Did they have any other concerns?

What suggestions do they have for making the questionnaire or individual items easier to understand and complete?
Pilot testing

- Obtain evidence of reliability.
- Establish evidence of face validity
- Obtain evidence of content validity
- Obtain evidence of criterion validity
- Obtain evidence of construct validity
Data Collection Methodology

- **Self-administered**
  - Individual, Letter
  - Group
  - Pooling
  - Email/internet

- **Observation**
  - Checklist

- **Combination of format and approach**
  - Practice+ Emotion
  - Checklist+ fill the blank+rating scales
Measurement

Validity

Reliability

Generalisibility
Validity and reliability

- Reliable but Not Valid
- Valid but Not Reliable
- Valid and Reliable
Apakah kita mengukur apa yang ingin kita ukur?

Konsep seringkali sulit diukur

- Misalnya:
  - Konsep : Pengetahuan.
  - Latent & Manifest Variable
Tipe Validity

- Face Validity
- Construct validity
- Content validity/internal validity
- Criterion validity
- Predictive validity
- Multicultural validity
Face validity is the degree to which an instrument appears to be an appropriate measure for obtaining the desired information, particularly from the perspective of a potential respondent.

- Smoking behavior → how many cigarettes a day they smoke → valid
- Healthy life style → how often do you exercise?
Construct Validity

- The degree to which an instrument measures an indirectly measurable concepts (construct), i.e. safety, intelligence, creativity, or patriotism.

- Ensuring that instrument designers and respondents have a shared definition of the construct.

- Related to the theoretical of knowledge

- May change over time

- Operationalization: the more factors or variables, we can associate with the concept the more valid our measurement will be.

- Involves:
  - Convergent validity: + depression and feelings of worthlessness.
  - Discriminant validity: - depression and feelings of happiness.
  - Demonstrate both
Convergent Validity

to show that measures that should be related are in reality related

<table>
<thead>
<tr>
<th></th>
<th>item 1</th>
<th>item 2</th>
<th>item 3</th>
<th>item 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>item 1</td>
<td>1.00</td>
<td>.83</td>
<td>.89</td>
<td>.91</td>
</tr>
<tr>
<td>item 2</td>
<td>.83</td>
<td>1.00</td>
<td>.85</td>
<td>.90</td>
</tr>
<tr>
<td>item 3</td>
<td>.89</td>
<td>.85</td>
<td>1.00</td>
<td>.86</td>
</tr>
<tr>
<td>item 4</td>
<td>.91</td>
<td>.90</td>
<td>.86</td>
<td>1.00</td>
</tr>
</tbody>
</table>

the correlations provide evidence that the items all converge on the same construct
Discriminant Validity

to show that measures that should *not* be related are in reality *not* related
Content/internal validity

- the degree to which an instrument is representative of the topic and process being investigated.
- the instrument should address the full range of specific topic/process typically identified by experts and discussed in research literature.
- Try to identify as many factors as possible that operationalize the construct.
Criterion Validity

- making a comparison between a measure and an external standard.

- Stroke recovery vs level of assistance required
  1. operationalize the concept of independent functioning by identifying activities of daily living
     - tying one’s shoes, getting dressed, brushing one’s teeth, and bed making,
  2. Compare to their actual performance
  3. Compare to results from another instrument that attempts to measure the same construct using the same criteria.
  4. If there is a strong relationship → criterion validity
Predictive validity

- to predict the results of one variable from another variable.

Example:
- Correlation of TOEFL Test to GPA
- Correlation of Psychometric test to staff loyalty
Multicultural validity

- An instrument measures what it purports to measure as understood by an audience of a particular culture.
- A multiculturally valid instrument will use language appropriate to its intended audience.
Demonstrating Validity: Qualitative

- Pretesting
- Qualitative
  - review research literature about the topic of interest.
  - Invite topic experts to review the instrument
  - Invite potential users to review the instrument
  - Identify poorly worded items
  - develop a table of specifications
    - Deductive
    - Inductive
Demonstrating Validity: Quantitative

- measuring the strength of the association between your instrument and another measure of the same construct.

- Convergent and discriminant validity

- Item analysis:
  - A valid item should be a good measure of what it is intended to measure and not of something else.

- Factor analysis
  - use correlations to identify common factors that influence a set of measures and individual factors that are unique to each item
Item Analysis

- Item analysis
  - To demonstrate a relationship between individual items
  - Internal consistency reliability
  - 1-2, 1-3, 1-4, 1-5, ...
  - 2-3, 2-4, 2-5, 2-6, ...
Exploratory Factor Analysis

- To identify the nature of constructs underlying responses in a specific content area.
- To determine what sets of items ‘hang together’ in a questionnaire, and
- To demonstrate the dimensionality of a measurement scale [where] researchers wish to develop scales that respond to a single characteristic.
Choose 10 top scorer and 10 lowest scorer
- Select randomly if there are more than 10 top/lowest scorer

Count how many subject in the top scorer group and lowest scorer group answer question 1 correctly, question 2, and so on..

Difficulty index: correct answers/total participants
- (RU+RL)/20

Discrimination index: (RU-RL)/10
- >0 : positive discrimination
- <0 : negative discrimination
<table>
<thead>
<tr>
<th>Name</th>
<th>Item 1</th>
<th>Difficulty Index: (8+4)/20 = 0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>Discrimination index (8-4)/10= 0.4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Compare to the maximum discriminating index</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>Near maximum: very discriminating</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>Half the maximum: moderately discriminating</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>A quarter the maximum: weak item</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>Near zero : non-discriminating</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>Negative: bad item</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>RU=8</td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>RL=4</td>
</tr>
</tbody>
</table>
Reliability

the extent to which an instrument produces the same information at a given time or over a period of time.
Source of random error

- Subject reliability: tired, moody
- Observer reliability: observer competence/interviewer, background
- Situational: situation where the interview or data collection take place (office / home)
- Instrument: bad wording
- Data processing: entry error, wrong coding
Establishing Evidence for Reliability

- Eyeballing: informal method,
  - administer the instrument twice to the same group of people in a relatively short period of time to see if their responses remain the same.

- Repeated measurement
  1. Test-retest method
     - **When?**
       - Carry-over effects
         - Too early: over-reliability
         - Too late: under-reliability
     - **How?**
       - Mengukur seberapa kuat hubungan score yang diukur pada 2 waktu yang berbeda dengan correlation coefficient
       - Reliable if coefficient correlation >0.7
Proportion agreement

- Use with discrete data (yes or no, male or female, and so on)
- The closer the number is to 1.0 the less likely it is that the results are due to random error.

Would you recommend this class to other students?  Yes  No

<table>
<thead>
<tr>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>b</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td>d</td>
</tr>
</tbody>
</table>

\[
\frac{a + d}{a + b + c + d} = \% \text{ of agreement.}
\]
If you want to obtain a measure of those who did not change their response over time, apply the values to the following formula (the symbol $\sum$ means the “sum of”; for example, $\sum a$ signifies all the $a$’s added together):

$$\frac{\sum a}{\sum a + \sum b + \sum c + \sum d + \sum e} = \% \text{ of agreement.}$$
Measure of association

- Pearson product moment correlation coefficient
  - >0.80 → strong correlations → stable

- Coefficient of determination: squared value of the correlation coefficient
  - $r$: 0.80
  - $r^2$: 0.64 → 64% due to variable of interest, 36% due to other factors
Inter-rater and Intra-rater Reliability

- Inter-rater: >1 rater
- Intra-rater: 1 rater
- Calculate with Cohen’s Kappa
Kappa Statistic (Cohen, 1960)

\[
\frac{OA}{EA} = \frac{OA}{1 - EA}
\]

OA: Kesepakatan yang terjadi
EA: Kesepakatan yg tidak disengaja

\[
OA = \frac{A+D}{N}
\]

\[
EA = \frac{N_1}{N} \cdot \frac{N_3}{N} + \frac{N_2}{N} \cdot \frac{N_4}{N}
\]

-1 < K < 1
## Agreement between observer 1 and 2

<table>
<thead>
<tr>
<th>Observer 2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>140</td>
<td>52</td>
<td>192</td>
</tr>
<tr>
<td>No</td>
<td>69</td>
<td>725</td>
<td>794</td>
</tr>
<tr>
<td>Total</td>
<td>209</td>
<td>777</td>
<td>986</td>
</tr>
</tbody>
</table>

### Observed agreement:

\[
\text{Observed agreement} = \frac{140 + 725}{986} = 0.877
\]

### Kappa:

\[
\text{Kappa} = \frac{0.877 - 0.676}{1.0 - 0.676} = 0.62
\]

### Chance agreements:

\[
\text{chance agreement between Yes - Yes} = \frac{209 \times 192}{986} = 40.7
\]

\[
\text{chance agreement between No - No} = \frac{777 \times 794}{986} = 625.7
\]

### Total expected change agreement:

\[
\text{total expected change agreement} = \frac{40.7 + 625.7}{986} = 0.676
\]
Test-Retest reliability

- pretest the questionnaire with the same group on two separate occasions, expecting only minor variations in responses.
- Coefficient of variation
- Similar to the eyeballing methods and proportion of agreement
Internal Consistency Reliability

- To compare results across and among items within a single instrument and to do so with only one administration.
- For multi-item scales
- Seberapa homogen item-item pertanyaan dalam 1 tes
- Seberapa baik item-item pertanyaan itu mengukur satu construct
- techniques
  - Average inter-item and average item-total correlation
  - split half reliability
  - coefficient alpha
  - Kuder Richardson
Average inter-item and average item total correlation
Internal Consistency Reliability

Split-half reliability
1. randomly split all the instrument items into two sets (even vs odd; first half vs second half)
2. ensuring that the items in the first set measure the same construct as the items in the second set do
3. Count the score for each set
4. Calculate coefficient correlations between set 1 and set 2
5. Reliable if coefficient correlation >0.8

Kuder-Richardson (KR)
- for estimating all possible split-half method correlations
- appropriate only for instruments intended to measure a single construct.
- applied only to instruments that use dichotomous items
Cronbach alpha

- To measure internal consistency
- Adopted from Kuder & Richardson (1937)
- Scaled/ranked data

- Internally consistent if coefficient alpha > 0.7
randomly split the items into two sets → compute the correlation between these sets → Put all the items back → randomly split them into two sets again → repeat for all possible split half correlations → calculate the average of all the correlations.

\[ \alpha = \frac{n}{n-1} \left( 1 - \frac{\sum V_i}{V_{test}} \right) \]

- Large Vtest → Small Ratio ΣVi/Vtest → high alpha
How alpha works

- \( V_i = p_i \times (1-p_i) \)
  - \( p_i = \) percentage of class who answers correctly
  - This formula can be derived from the standard definition of variance.

- \( V_i \) varies from 0 to 0.25

<table>
<thead>
<tr>
<th>( p_i )</th>
<th>1-( p_i )</th>
<th>( V_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.25</td>
<td>0.75</td>
<td>0.1875</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>
What if instrument is not reliable?

- Identify a not reliable question(s)
- Identify strength of correlation between item and total score
- Low correlation item will reduce instrument’s reliability and better be removed.
- In test and retest method, look at question which has a big gap score between test and retest.
How to improve reliability?

- Make sure that the questions are clear and not ambiguous
- Make it specific
- Create several questions to measure one construct, but not too many
From sample to population

Sample: true exist or just a coincidence

<table>
<thead>
<tr>
<th>Your decision (based on the sample)</th>
<th>The real situation (in the population)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ is true</td>
</tr>
<tr>
<td>$H_0$ is supported</td>
<td>No error</td>
</tr>
<tr>
<td>$H_1$ is supported</td>
<td>Type I error</td>
</tr>
</tbody>
</table>
Is there a relationship between student’s healthy lifestyle and healthy school program?

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Result</th>
<th>Reality</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis (H₀): No relationship between clean lifestyle and healthy school program.</td>
<td><strong>There is a relationship</strong> between clean lifestyle and healthy school program.</td>
<td>Null hypothesis rejected</td>
<td></td>
</tr>
<tr>
<td>Alternative hypothesis (H₁): There is a relationship between clean lifestyle and healthy school program.</td>
<td><strong>No relationship</strong> between clean lifestyle and healthy school program.</td>
<td>Type 1 error</td>
<td></td>
</tr>
</tbody>
</table>

Implication: more healthy school program
<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Hasil penelitian</th>
<th>Kenyataan di populasi</th>
<th>Interpretasi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis ((H_0)): No relationship between clean lifestyle and healthy school program.</td>
<td>No relationship between clean lifestyle and healthy school program.</td>
<td><strong>Null hypothesis accepted</strong></td>
<td></td>
</tr>
<tr>
<td>Alternative hypothesis ((H_1)): There is a relationship between clean lifestyle and healthy school program.</td>
<td>There is a relationship between clean lifestyle and healthy school program.</td>
<td></td>
<td><strong>Type 2 error</strong> Reduced healthy school program</td>
</tr>
</tbody>
</table>
How big is the chance for type 1 error?

- Measure with level of significance / p-values / coefficient alpha
- Smaller coefficient alpha $\rightarrow$ smaller chance of type 1 error
- Common cut-off point is $p<0.05 \rightarrow$ significant
- Influenced by:
  - sample size
  - Variation within sample
- Interpretation
  - How do you interpret $p=0.052$ and $p=0.049$?
If relationship between two variable show p < 0.05, does it mean important finding?

If effect size between variable is big, does it mean the relationship is important?

Is internal consistency reliability and construct validity measure the same thing?

If statistics analysis showing a significant result does it mean the phenomenon could be find in the general population?
Quality in Qualitative Data

Accurate
- recall, transcription, interpretation,

Contexted
- setting, social context, body language, tone, feeling,

Thick description

Useful

Reflexive
- it’s you and your data
Quality of Qualitative Data

Trustworthiness

- Credible
- Dependable
- Confirmable
- Transferable
## Credibility (validity)

<table>
<thead>
<tr>
<th>Truth value or confidence in the truth of the findings</th>
<th>Do the findings show a logical relationship to each other?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Are they consistent in terms of the explanations they support?</td>
</tr>
<tr>
<td></td>
<td>Are the narrative data sufficiently rich to support the specific findings?</td>
</tr>
<tr>
<td></td>
<td>Do the findings indicate a need for more data?</td>
</tr>
<tr>
<td></td>
<td>Does the original study population consider reports to be accurate?</td>
</tr>
</tbody>
</table>
Dependability (reliability)

The research process is consistent and carried out with careful attention to the rules and conventions of qualitative methodology.

- Are the research questions clear and logically connected to the research purpose and design?
- Are there parallels across data sources?
- Do multiple field-workers have comparable data collection protocols?
Confirmability (objectivity)

- minimizing any possible influence of the researcher’s values on the process of inquiry.
- maintained the distinction between personal values and those of the study participants.
- Reflexivity
Transferability (generalizability)

- whether the conclusions of a study are transferable to other contexts
- The researcher must account for contextual factors when transferring data from one situation to another.
How to improve trustworthiness

- Thick description
- Negative/defiant case analysis
- Triangulation (data, subject, methods)
- Member checking