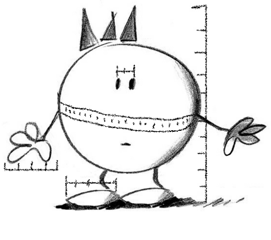
Research methods 09

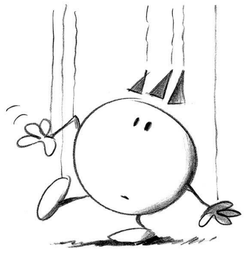
Measurement of Constructs
Reliability

Caryn Block
ORLJ 5040
Teachers College
Columbia University

Operationalizing variables

Measurement vs. **Manipulation**





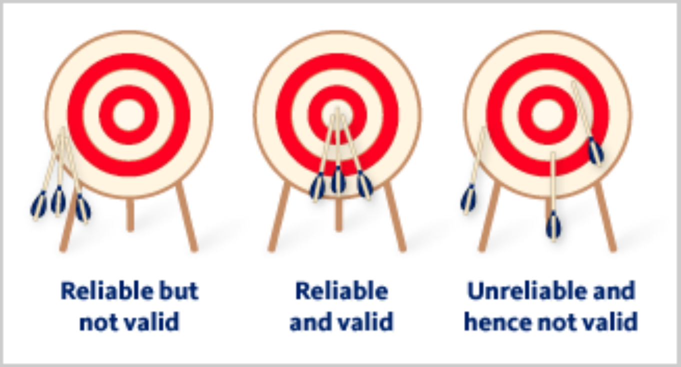
- assigning numbers to people

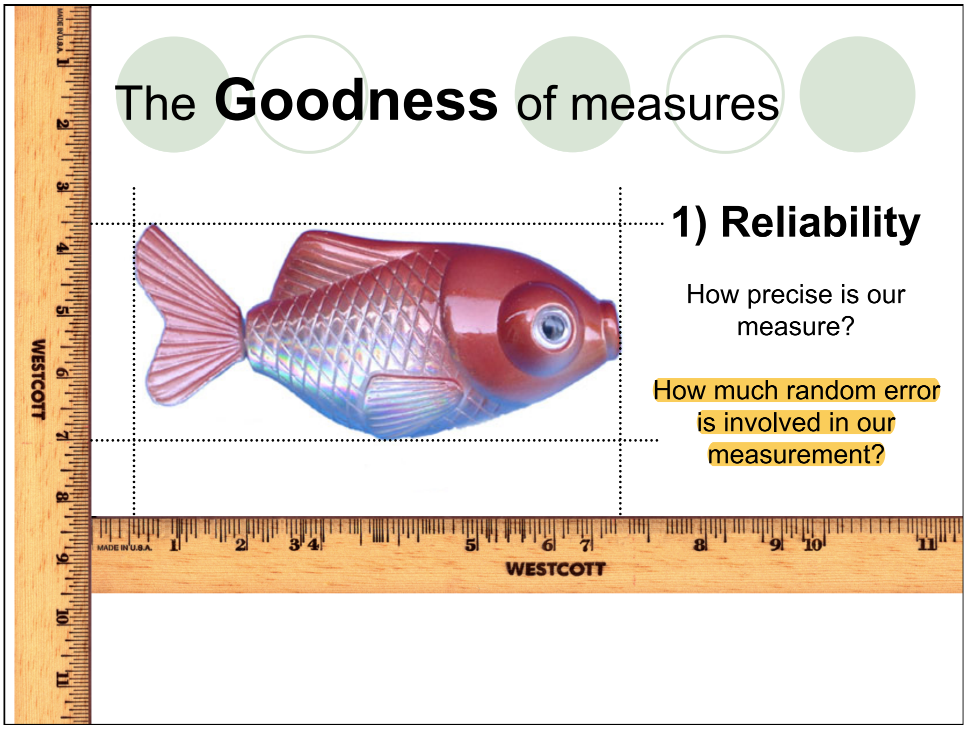
- changing people’ s experience and behavior in a systematic way

Reliability and Validity

- Two most important psychometric characteristics of a measure
- Answers two different questions:
  - **Reliability:** Is a test/assessment dependable, stable, and/or consistent over time?
  - **Validity:** Does a test/assessment measure what it is supposed to measure?
- Reliability is a necessary, but insufficient condition for validity (APA, 1995)
- No cut-off criterion of establishing reliability & validity, it’ s more about accumulation of evidence that suggests that the test measures what it purports to measure

Reliability and Validity



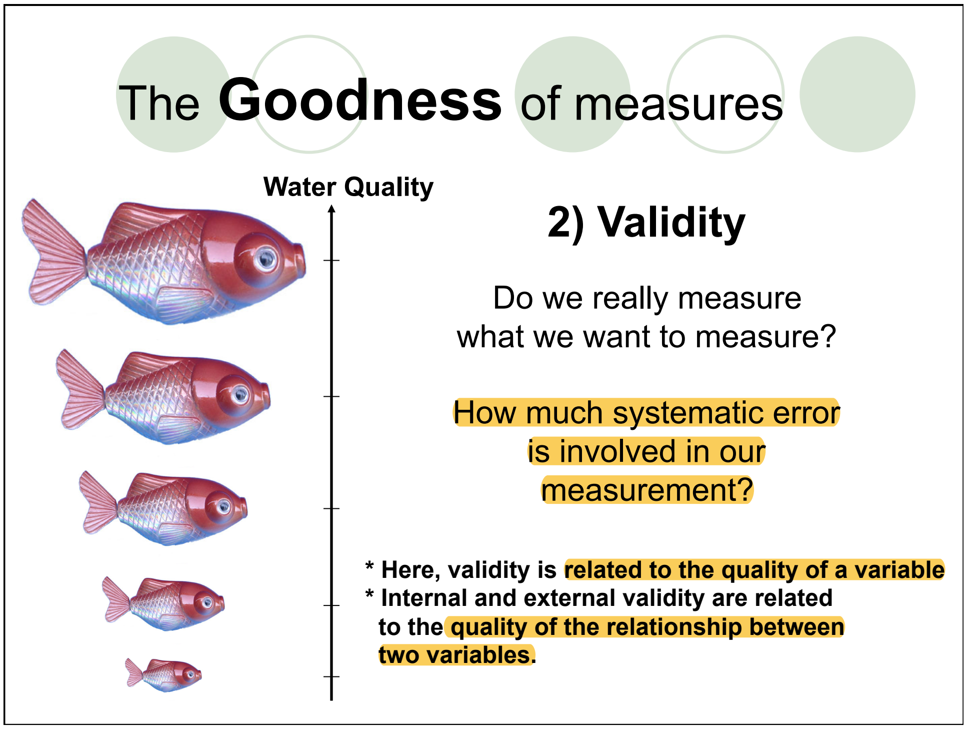


The **Goodness** of measures

1) Reliability

How precise is our measure?

How much random error is involved in our measurement?



The **Goodness** of measures

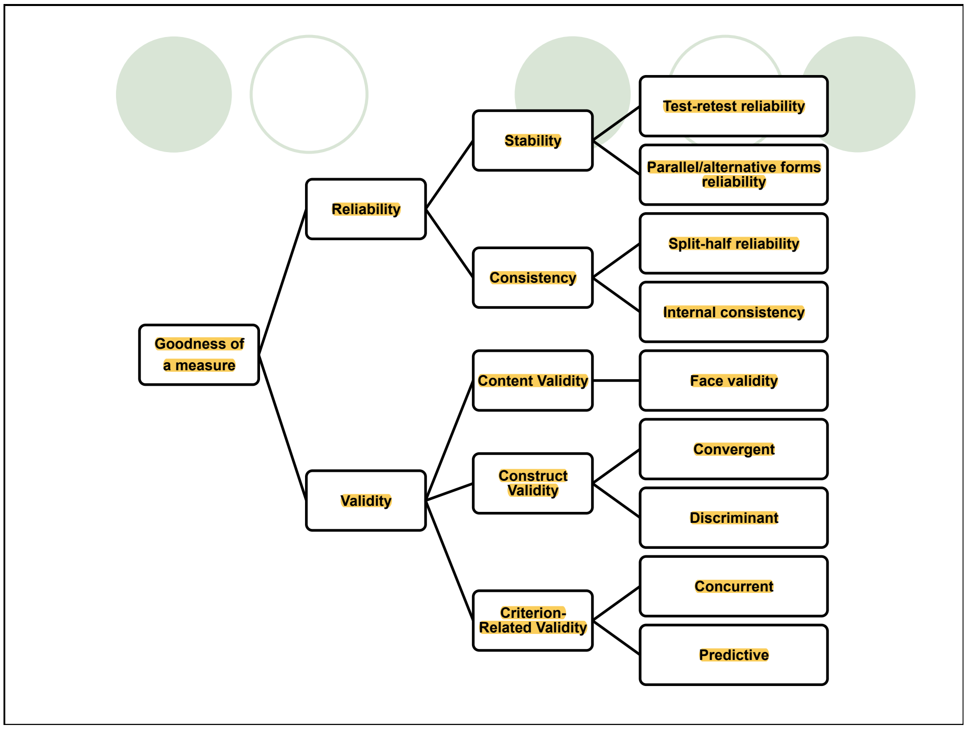
2) Validity

Do we really measure what we want to measure?

How much systematic error is involved in our measurement?

*** Here, validity is related to the quality of a variable**

*** Internal and external validity are related to the quality of the relationship between two variables.**



Goodness of a measure

Reliability

Stability

Test-retest reliability

Parallel/alternative forms reliability

Consistency

Split-half reliability

Internal consistency

Validity

Content Validity

Face validity

Construct Validity

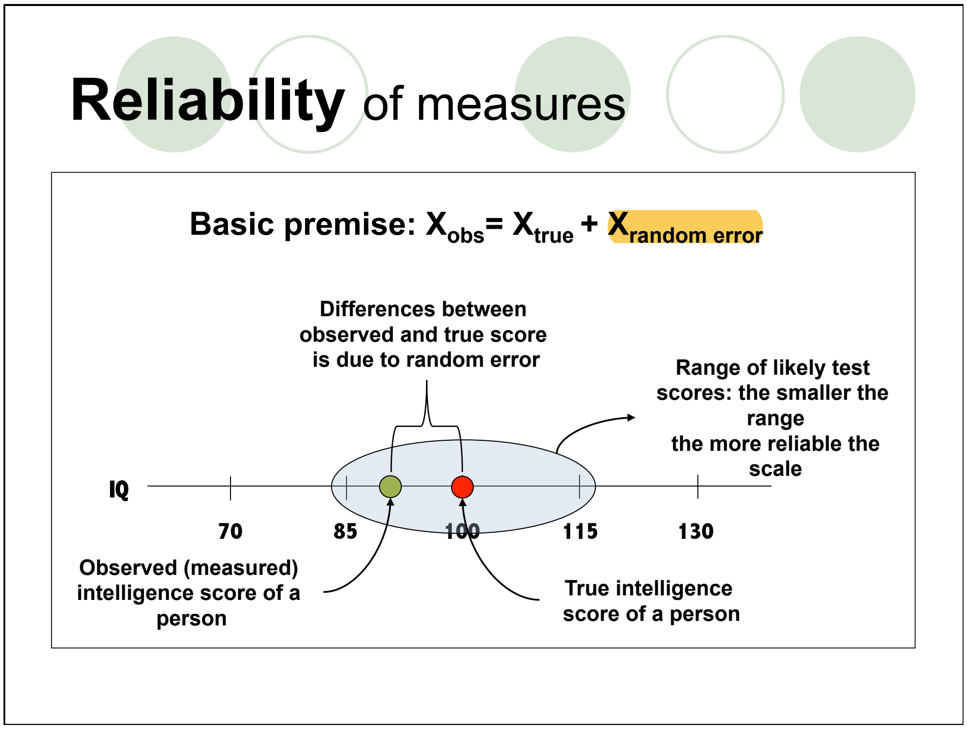
Convergent

Discriminant

Criterion-Related Validity

Concurrent

Predictive



Reliability of measures

Basic premise: $X_{obs} = X_{true} + X_{random\ error}$

Differences between observed and true score is due to random error

Range of likely test scores: the smaller the range the more reliable the scale

Observed (measured) intelligence score of a person

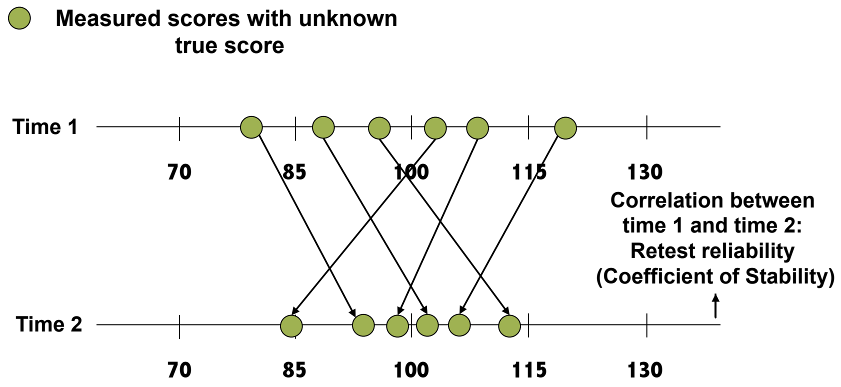
True intelligence score of a person

Estimating Reliability

- True score is the score an individual would obtain if all internal and external conditions were perfect and the measuring instrument was perfect
- The reliability of a variable cannot be measured directly because the true scores are unknown
- Random error is unrelated between measurements
- Therefore, true scores can be estimated by the average of multiple measurements:
  - ○ X true = (X1 + X2 + X3 + …+ Xn)/n
- Random error is ‘washed out’ by averaging process
- Different measurements of same score can be
  - ○ Same variable measured several times
  - ○ Different variables (items) that measure the same construct

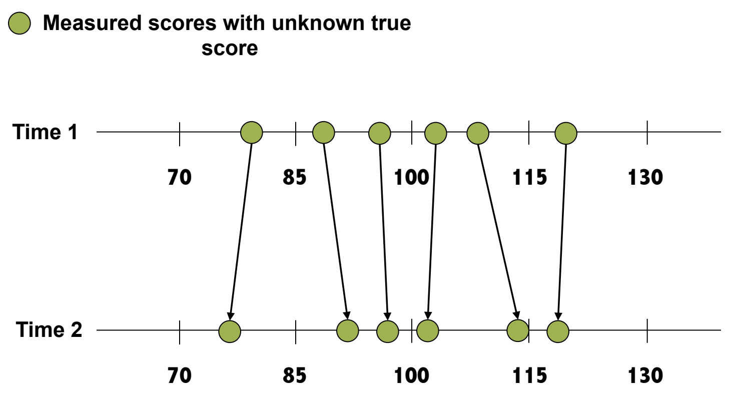
Estimating Reliability: **Retest Reliability**

stability


● Measured scores with unknown true score
Time 1
70
85
100
115
130
Time 2
70
85
100
115
130
Correlation between time 1 and time 2: Retest reliability (Coefficient of Stability)

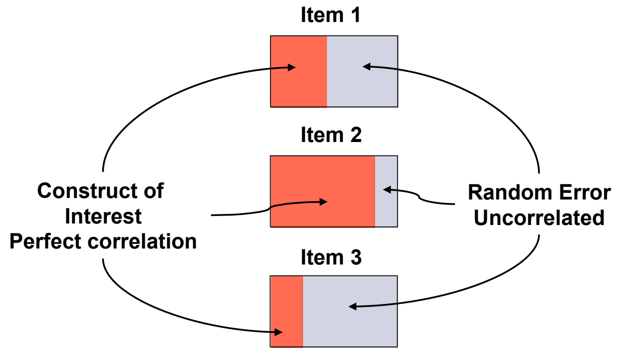
The more scores change over time, the less reliable the test

Estimating Reliability: **Retest Reliability**



The less scores change over time, the more reliable the test

Estimating Reliability: **Internal Consistency**



- The more items of a measure are correlated with each other, the higher is the reliability of that measure.
- The less items are correlated, the more random error is involved in measuring the construct of interest.

Internal Consistency:
Split-half Reliability

1. Administer test
2. Split test in half (e.g., even-odd number split)
3. Calculate the correlation between the two halves

Subject	Odd half	Even half
1	47	46
2	20	19
3	36	47
4	45	39
5	25	28

r = .86

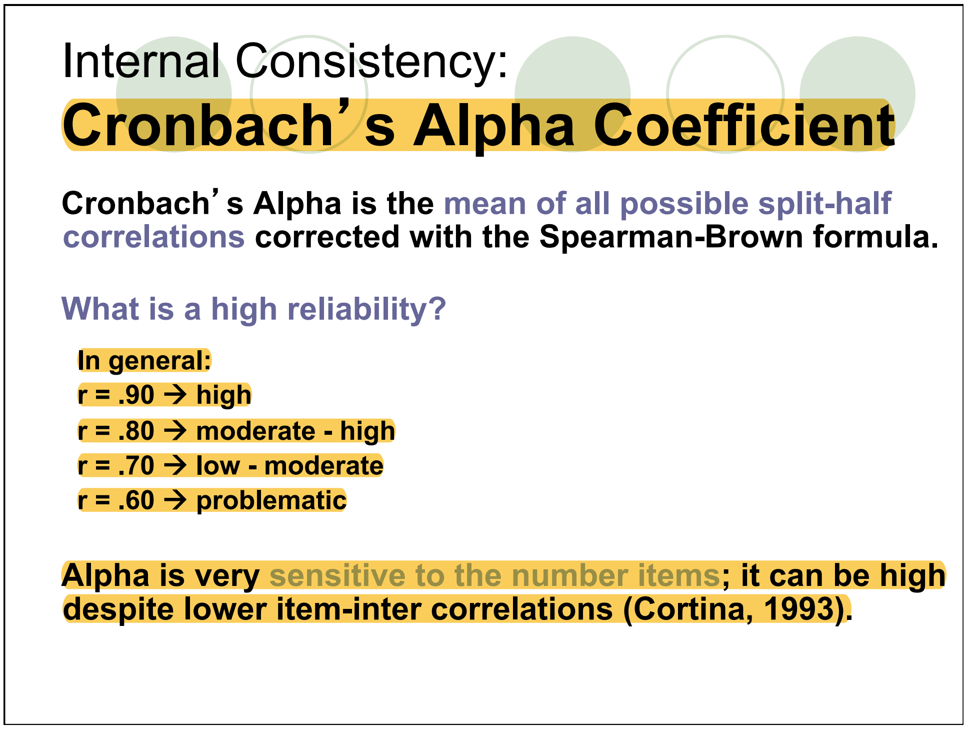
Internal Consistency:
Split-half Reliability

Problem with split-half reliability:
A shorter measure is tested (2 measures with half the items)

- The less items the lower the correlation
- Brown-Spearman correction formula:

$$r_{SB} = \frac{nr}{1+(n-1)r}$$

- r - split-half correlation
- n - number of total items



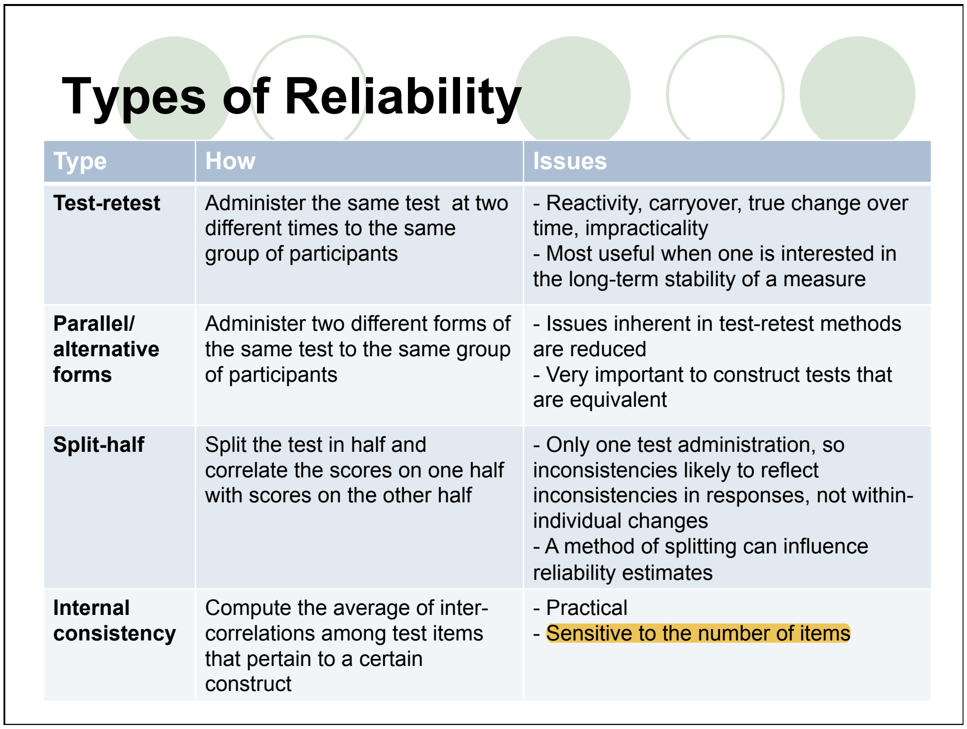
Internal Consistency:
Cronbach's Alpha Coefficient

Cronbach's Alpha is the mean of all possible split-half correlations corrected with the Spearman-Brown formula.

What is a high reliability?

In general:
r = .90 → high
r = .80 → moderate - high
r = .70 → low - moderate
r = .60 → problematic

Alpha is very sensitive to the number items; it can be high despite lower item-inter correlations (Cortina, 1993).



Types of Reliability

Type	How	Issues
Test-retest	Administer the same test at two different times to the same group of participants	- Reactivity, carryover, true change over time, impracticality - Most useful when one is interested in the long-term stability of a measure
Parallel/ alternative forms	Administer two different forms of the same test to the same group of participants	- Issues inherent in test-retest methods are reduced - Very important to construct tests that are equivalent
Split-half	Split the test in half and correlate the scores on one half with scores on the other half	- Only one test administration, so inconsistencies likely to reflect inconsistencies in responses, not within-individual changes - A method of splitting can influence reliability estimates
Internal consistency	Compute the average of inter-correlations among test items that pertain to a certain construct	- Practical - Sensitive to the number of items

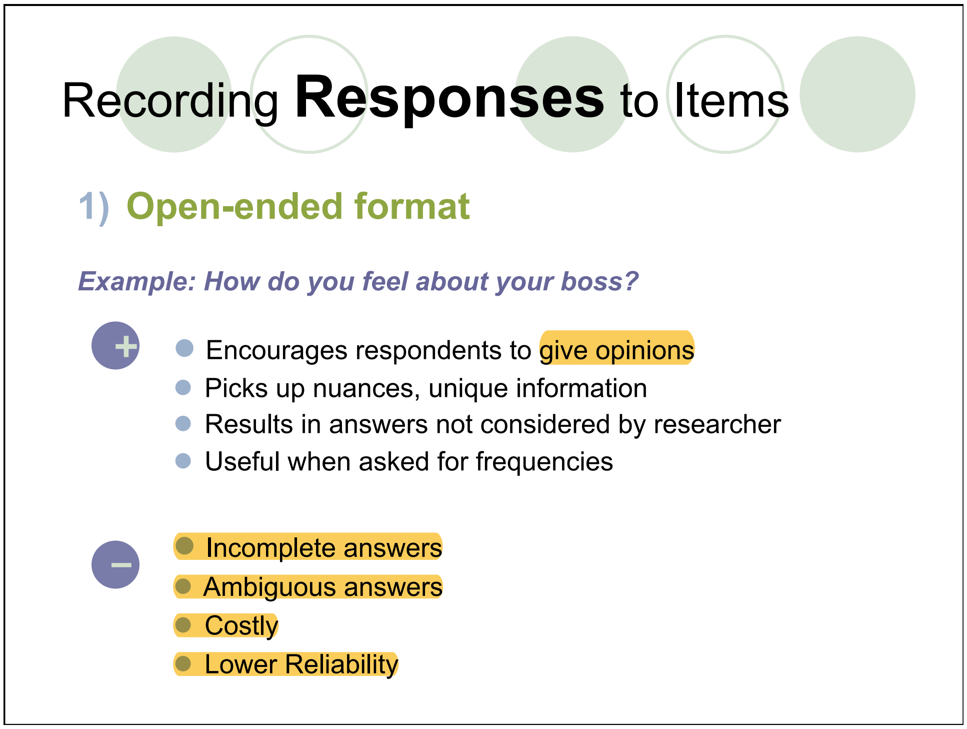
How to **increase** reliability?

- Longer measures are more reliable than shorter ones.
- The more variability among individuals the higher the reliability.
- Freedom from distractions and misunderstandings:
  - Clear instructions
  - Optimal test setting
- Clear and unambiguous items

Item Wording

Problems in Writing Items

- Ambiguity
- Jargon
- Length
- Double-barreled
- Leading
- Loaded
- Threatening
- Over-demanding
- Over-specificity
- Relevance



Recording **Responses** to Items

1) Open-ended format

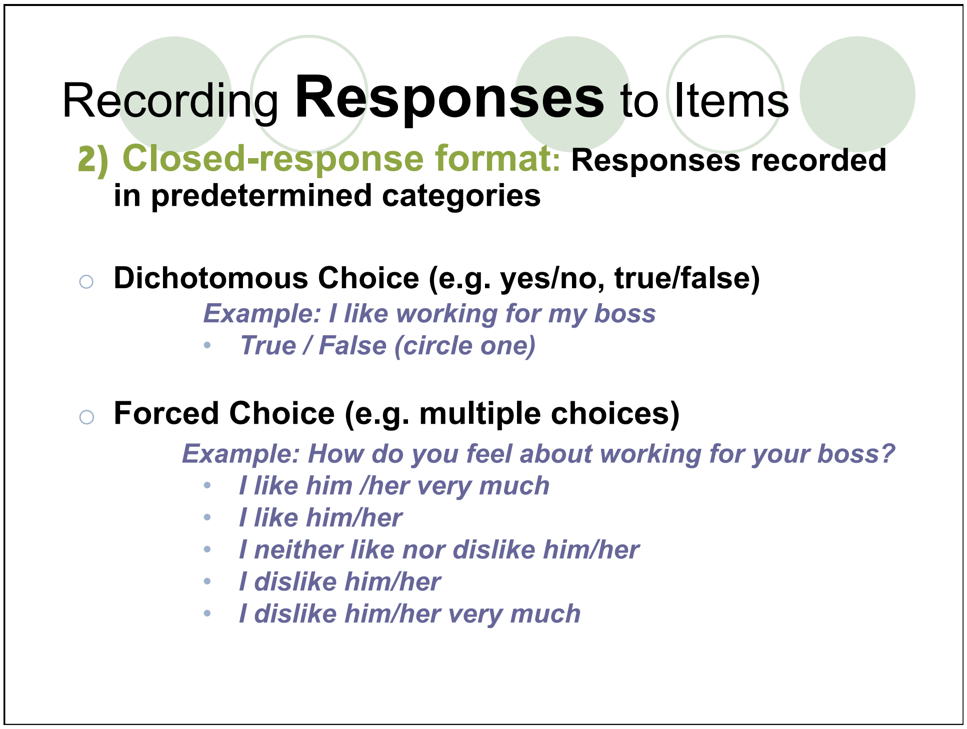
Example: How do you feel about your boss?

+

- Encourages respondents to give opinions
- Picks up nuances, unique information
- Results in answers not considered by researcher
- Useful when asked for frequencies

−

- Incomplete answers
- Ambiguous answers
- Costly
- Lower Reliability



Recording **Responses** to Items

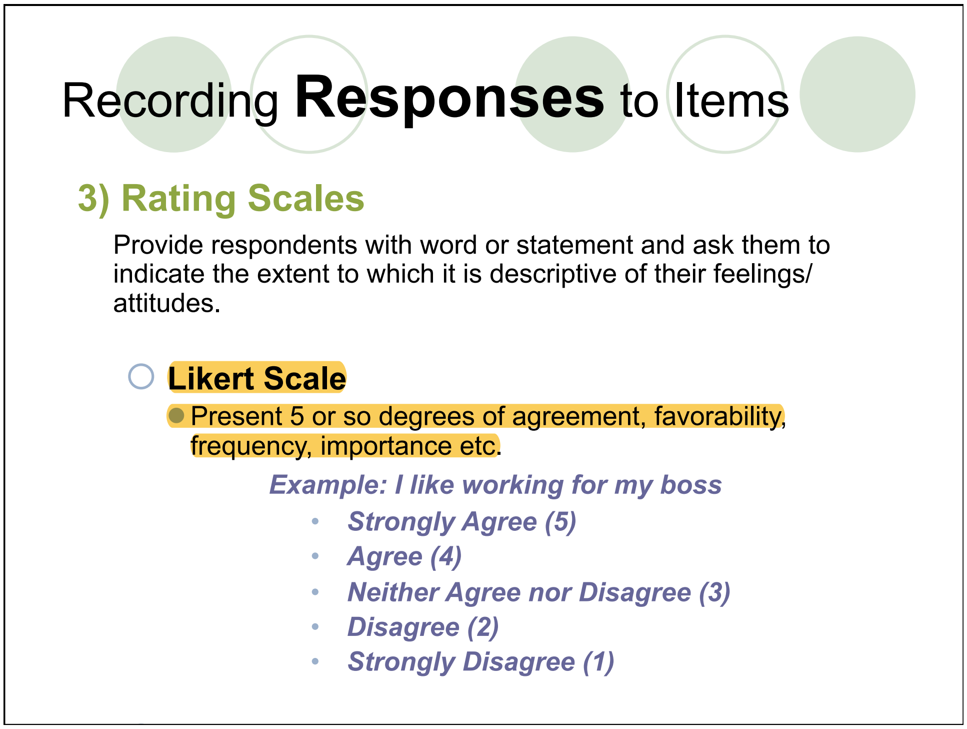
2) Closed-response format: Responses recorded in predetermined categories

- **Dichotomous Choice (e.g. yes/no, true/false)**

  Example: I like working for my boss
  - *True / False (circle one)*

- **Forced Choice (e.g. multiple choices)**

  Example: How do you feel about working for your boss?
  - *I like him /her very much*
  - *I like him/her*
  - *I neither like nor dislike him/her*
  - *I dislike him/her*
  - *I dislike him/her very much*



Recording **Responses** to Items

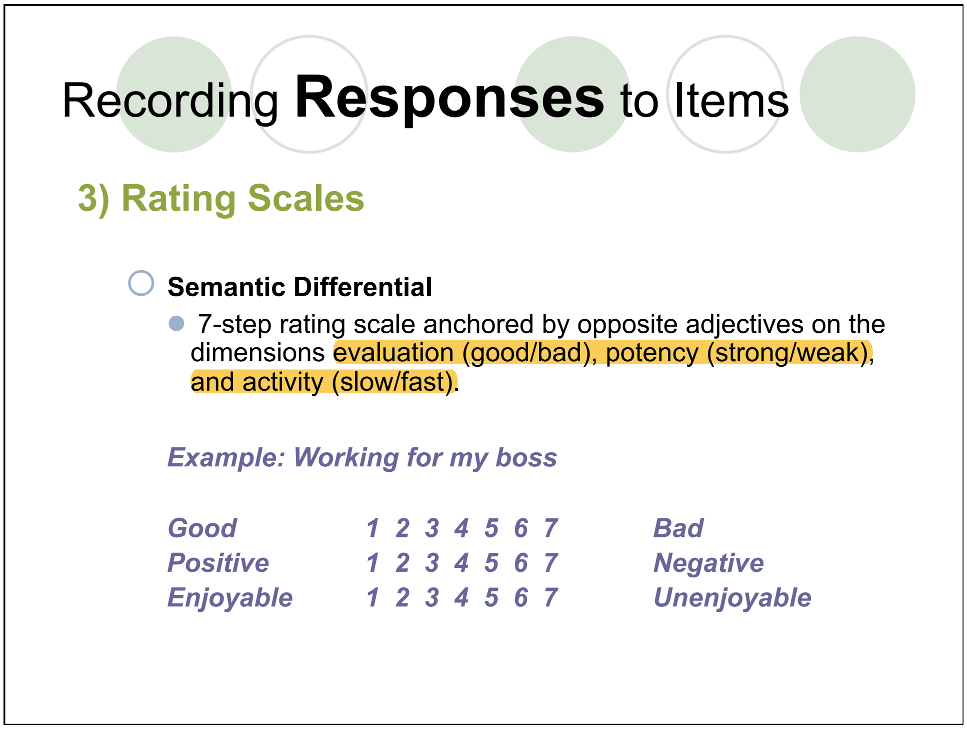
3) Rating Scales

Provide respondents with word or statement and ask them to indicate the extent to which it is descriptive of their feelings/ attitudes.

- ○ **Likert Scale**
  - Present 5 or so degrees of agreement, favorability, frequency, importance etc.

    Example: I like working for my boss

    - ***Strongly Agree (5)***
    - ***Agree (4)***
    - ***Neither Agree nor Disagree (3)***
    - ***Disagree (2)***
    - ***Strongly Disagree (1)***



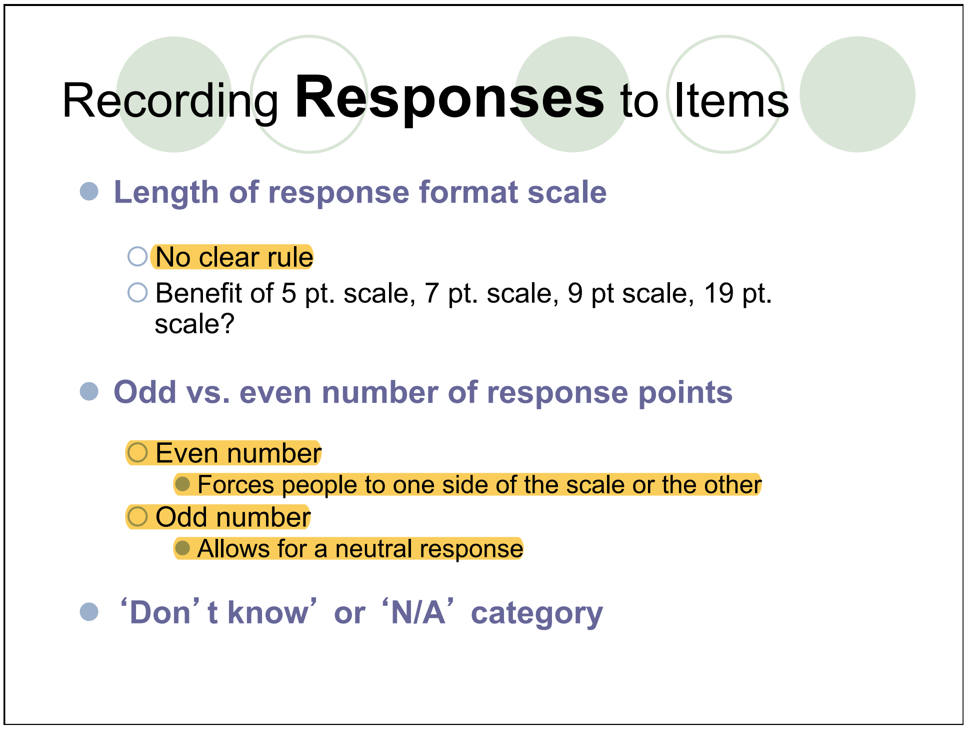
Recording **Responses** to Items

3) Rating Scales

- ○ **Semantic Differential**
  - 7-step rating scale anchored by opposite adjectives on the dimensions evaluation (good/bad), potency (strong/weak), and activity (slow/fast).

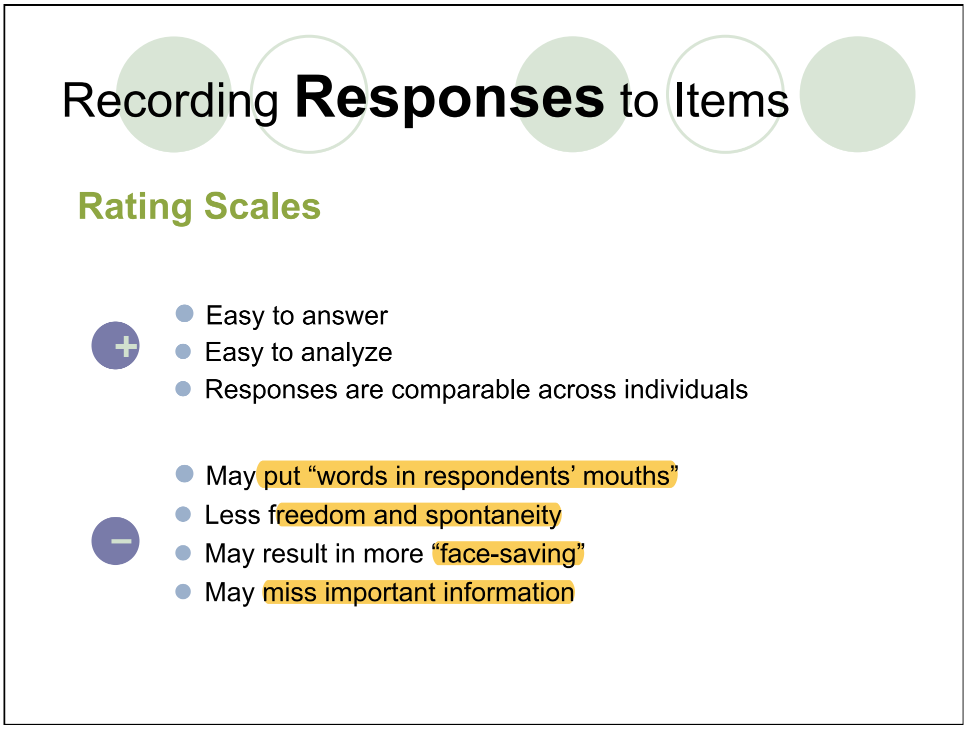
Example: Working for my boss

Good	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***Bad***
Positive	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***Negative***
Enjoyable	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***Unenjoyable***



Recording **Responses** to Items

- **Length of response format scale**
  - No clear rule
  - Benefit of 5 pt. scale, 7 pt. scale, 9 pt scale, 19 pt. scale?
- **Odd vs. even number of response points**
  - Even number
    - Forces people to one side of the scale or the other
  - Odd number
    - Allows for a neutral response
- **'Don't know' or 'N/A' category**



Recording **Responses** to Items

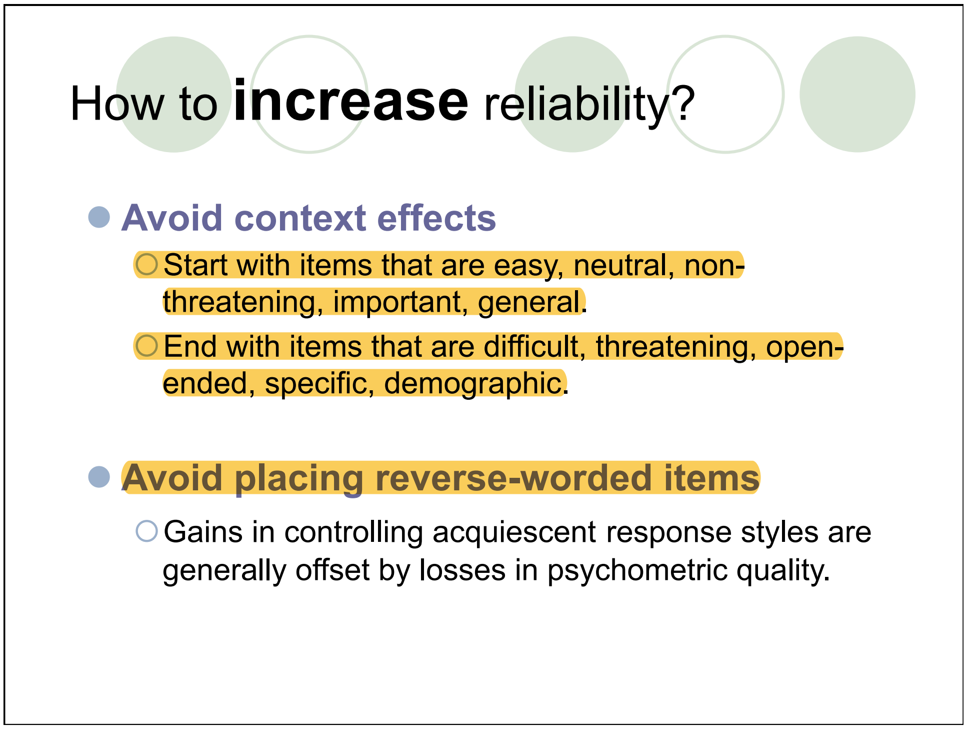
Rating Scales

+

- Easy to answer
- Easy to analyze
- Responses are comparable across individuals

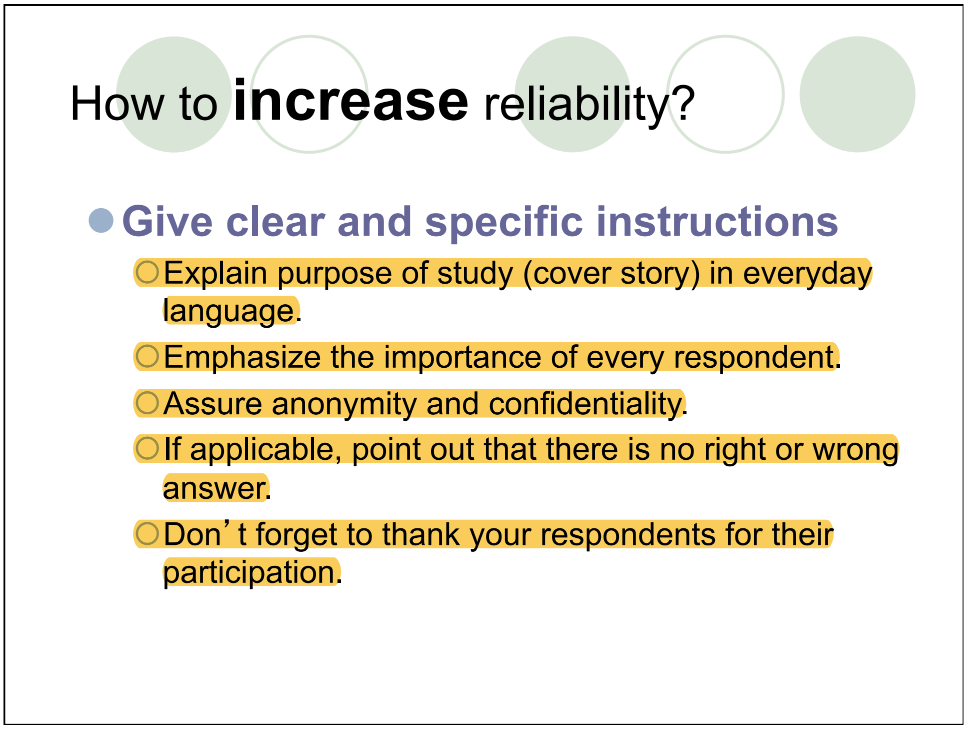
−

- May put "words in respondents' mouths"
- Less freedom and spontaneity
- May result in more "face-saving"
- May miss important information



How to **increase** reliability?

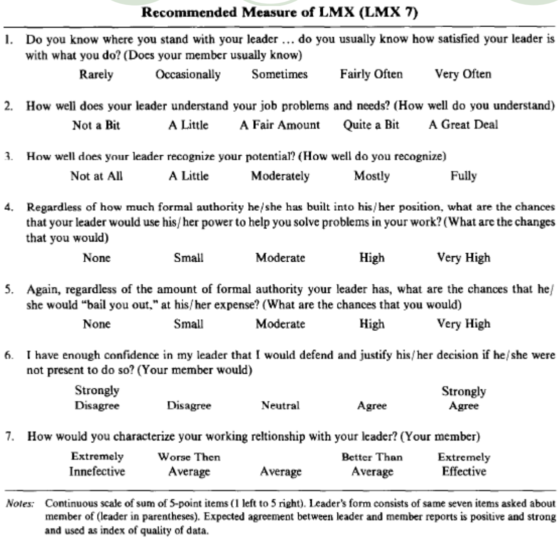
- **Avoid context effects**
  - Start with items that are easy, neutral, non-threatening, important, general.
  - End with items that are difficult, threatening, open-ended, specific, demographic.
- **Avoid placing reverse-worded items**
  - Gains in controlling acquiescent response styles are generally offset by losses in psychometric quality.



How to **increase** reliability?

- **Give clear and specific instructions**
  - Explain purpose of study (cover story) in everyday language.
  - Emphasize the importance of every respondent.
  - Assure anonymity and confidentiality.
  - If applicable, point out that there is no right or wrong answer.
  - Don’ t forget to thank your respondents for their participation.

LMX 7 Scale



Recommended Measure of LMX (LMX 7)

1. Do you know where you stand with your leader … do you usually know how satisfied your leader is with what you do? (Does your member usually know)

   Rarely Occasionally Sometimes Fairly Often Very Often

2. How well does your leader understand your job problems and needs? (How well do you understand)

   Not a Bit A Little A Fair Amount Quite a Bit A Great Deal

3. How well does your leader recognize your potential? (How well do you recognize)

   Not at All A Little Moderately Mostly Fully

4. Regardless of how much formal authority he/she has built into his/her position, what are the chances that your leader would use his/her power to help you solve problems in your work? (What are the changes that you would)

   None Small Moderate High Very High

5. Again, regardless of the amount of formal authority your leader has, what are the chances that he/she would "bail you out," at his/her expense? (What are the chances that you would)

   None Small Moderate High Very High

6. I have enough confidence in my leader that I would defend and justify his/her decision if he/she were not present to do so? (Your member would)

   Strongly Disagree Disagree Neutral Agree Strongly Agree

7. How would you characterize your working reltionship with your leader? (Your member)

   Extremely Innefective Worse Then Average Average Better Than Average Extremely Effective

Notes: Continuous scale of sum of 5-point items (1 left to 5 right). Leader's form consists of same seven items asked about member of (leader in parentheses). Expected agreement between leader and member reports is positive and strong and used as index of quality of data.

- What are the issues with the scale?
- How would you improve the scale?

Guidelines for Question Writing

- ✓ Relevance to research and respondents, ease of coding and cognitive demands placed on respondents
- ✓ Question sequencing: related questions together; sensitive items later; opening questions easy; use of funnel technique
- ✓ Wording should be simple, direct and familiar to all respondents. Consider respondents' writing and reading level and frame of reference
- ✓ Questions should be as clear and unambiguous as possible
- ✓ Questions should be applicable to all respondents
- ✓ Avoid double-barreled questions
- ✓ Avoid leading and loaded questions
- ✓ Minimize the influence of response styles

From Professor Elissa Perry's class: Understanding Behavioral Research ORLJ 4009