

Ch 11 – Self Report Measurement

I. Introductory Comments

- Purpose of measurement: make social phenomenon or theoretical constructs visible in a numerical format
- Operational definitions of variables include procedures for measuring the phenomena of interest.
 - E.g. Type of scale and how it will be interpreted.

- Focus on Psychometric Scaling (as opposed to psychophysical scaling).

Psychometric Scaling:

- developed in early portion of the 20th century.
 - Mostly in the 20s and 30s
- measure mental phenomenon with no clearly identifiable external cause, but the experience can be reported by the individual
 - (e.g. feelings, beliefs, Intelligence and other mental processes).
- Psychophysical Scaling = external cause

II. Open Ended Vs. Closed Ended Scales

- When self administered, use Closed-Ended scales
 - open ended = too little information, irrelevant information, or vague/uncodable information.
- Use Open ended when:
 - 1) Asking about socially undesirable behavior (people under-report negative behaviors on closed ended questions).
 - 2) Unsure what options to include (e.g. preliminary research)
 - Treat participants as SMEs.

III. Closed Ended Scales

1. Response Formats

- a) Comparative Rating Scales
- b) Itemized Rating Scales
- c) Graphical Rating Scales
- d) Numerical Rating Scales
- e) Unipolar vs. Bipolar Scales

a) Comparative Rating Scales (Provide ordinal data)

a1. Paired Comparisons

-compare item pairs in a single dimension

-e.g. a gopher and a weasel on attractiveness

- e.g. gopher vs. weasel

gopher vs. ferret

weasel vs. ferret

-Score = number times item is top of pair.

Prob. as # of items to be compared increase the number of comparisons rapidly increases; $[n(n-1)/2]$
10 items requires 45 comparisons.

a2. Rank ordering

- order all items along some dimension
- e.g. a gopher, weasel, ferret, ground hog, and chip monk on attractiveness
- Score is the rank received.

Rules:

- 1) Respondents must be familiar with all the stimuli.
- 2) Rating dimension must be Unidimensional.
- 3) Respondent must understand the dimension

a2. Rank ordering cont.

Problems

- Large Sets (more than 20) :
 - difficult to order
 - middle items tend to have arbitrary orders.
(Creating end-point pairs, best/worst, until list is exhausted can help).
- procedure may produce an artificial hierarchy; no ties.
- does not say where an item falls on the dimension (only gives relative location)
- Ranks are ordinal data – require non-parametric analyses.

b) Itemized Rating Scales (Provides Nominal, Ordinal, or Interval Data)

b1. Classifications (Nominal)

- Pick the mutually exclusive category that is most like you or your behavior/attitudes.

1. With which of the following small woodland creatures would you prefer to be trapped on a desert island:

- A) gopher B) weasel C) ferret
D) ground hog E) chip monk F) badger

- Categories: theory based or derived from open-ended research.
- Forced choice format can't identify ties between categories
- Assumes category is equally representative of all people who choose it.

b2. Ordinal

- Use same classification for multiple items (e.g. different situations)

-score = the number of times a specific classification was chosen

2. Of the following small woodland creatures, which would you prefer to give you a Swedish massage:

- A) gopher B) weasel C) Ferret
D) ground hog E) chip monk F) badger

Considering item 1 and 2 together will produce summative scores (sum of 1 & 2) for all six animal categories.

- Also classifications may represent an ordinal continuum.

3. Compared to your peers, how strong is your attraction to small woodland creatures?

- Very weak
- Weak
- Moderately weak
- Moderately strong
- Strong
- Very Strong

c. Graphical Rating Scales

Ordinal Level Measures

- Marking a line anchored by extremes of the dimension.
- The Feelings Thermometer.
- Segmented Line system

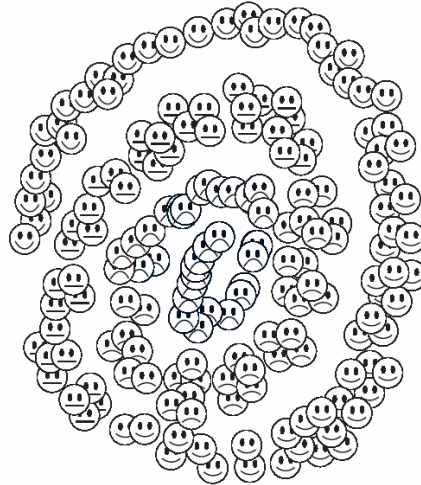
e.g. _____ , _____ , _____ , _____ , _____
= a 5 pt rating system

- For children

- The Smiling, Frowning, and Ambivalent faces.
- E.g. Badgers make me feel:



The far less popular Spiral of Doom rating scale



d. Numerical Rating Scales

- Circle, mark, or write numbers that are paired with dimensional anchors.
- Probably most common system.

- E.g., Putting a badger in your pants is perfectly acceptable.

(1)	(2)	(3)	(4)	(5)
Strongly Disagree	Disagree	Neither or Both	Agree	Strongly Agree

- Could also have people provide values

- Frequency =
 - What percent of the day do you spend thinking about putting badgers in your pants _____.

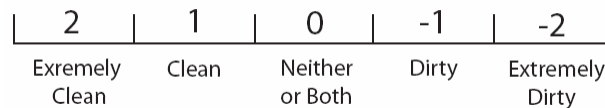
- The Size of Range
 - Sensitivity of Measurement.
 - more points on the scale will give more sensitive measure of the dimension.
 - Usability
 - Too many points and too few points make it difficult for people to decide how to answer.
 - Ideal number of points is 5 - 9.

- Also, scales with less than 4 points tend to behave like ordinal data rather than interval.

- Anchors.
 - Anchors should be equal interval
- Agreement Agree/Disagree
- Evaluation Excellent - Terrible
 Good - Bad
 Like – Dislike
- Frequency Always – Never
- Amount All – None
- Similarity Very Much Like Me
 Not at all Like Me

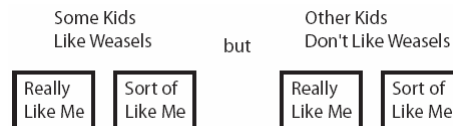
- Equal-Appearing Intervals.
- Combination of Graphical Scale and a Numerical Rating Scale
- Used in Thurstone scaling
- Response points are presented such that they appear equidistant.

Putting Badgers in Your Pants is:



Kids Formats

- younger children can have difficulty with standard numerical rating scales
- Kids - Some Kids / Other Kids Format
- Some Kids like weasels, but other Kids do not. Which kids are more like you?



- Basically results in a 4 point scale

e. Unipolar vs. Bipolar Scales (Fishbein & Ajzen, 1975).

- Unipolar

Hot - Not Hot

Cold - Not Cold

Dirty - Not Dirty

Clean - Not Clean

0	1	2	3
Not at all	Slightly	Quite	Extremely
Hot	Hot	Hot	Hot

0	1	2	3
Not at all	Slightly	Quite	Extremely
Cold	Cold	Cold	Cold

- Bipolar/Bidirectional Scale

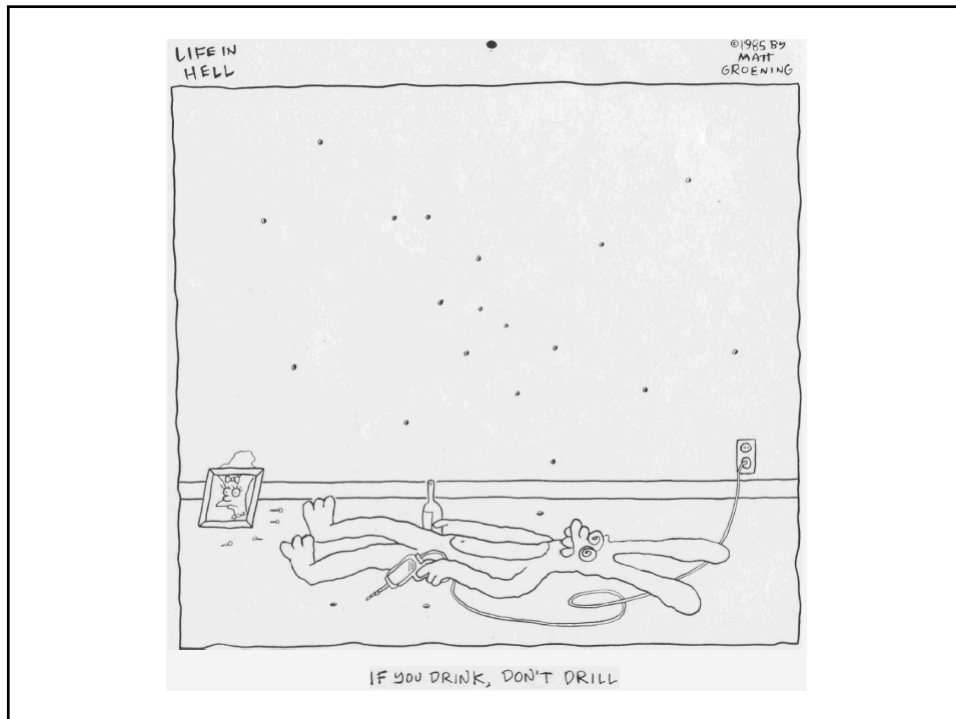
Hot - Cold

Dirty - Clean

E.g.

(3)	(2)	(1)	(0)	(-1)	(-2)	(-3)
Extremely Clean	Quite Clean	Slightly Clean	Neither or Both	Slightly Dirty	Quite Dirty	Extremely Dirty

- In practice, responses obtained from two unidirectional scales tend to be correlated with scores on a single bi-directional scale.
- However, it is theoretically problematic to try to predict scores on unidirectional scales from scores on a single bi-directional scale.
 - E.g. -1 on a bidirectional scale could represent the following
 - a. quite clean and extremely dirty
 - b. slightly clean and quite dirty
 - c. not at all clean and slightly dirty



IV. Closed Ended Multi-Item Scales

A. **Guttman Scaling** (Cumulative Scales): scale items are ordered such that answering affirmative to an item presupposes that previous items were answered yes to (Unidimensional).

Drinking and Drilling

- | | | |
|--|---|---|
| 1. Do you drink | Y | N |
| 2. Do you drink & drill | Y | N |
| 3. Have you ever had an accident while drinking & drilling | Y | N |

Scale score is equal to the question item that you stopped answering yes (or no) to.

Responses can be arranged into a Scalogram.

Below is a perfect unidimensional scale

subject	Q1	Q2	Q3	Score
1	0	0	0	0
2	1	0	0	1
3	1	1	0	2
4	1	1	1	3

- Coefficient of Reproducibility

- Degree to which Scalogram deviates from perfect unidimensional matrix.

- E = number of errors (inappropriate responses)

$$R = 1 - \frac{E}{n(k)}$$

- n = number of respondents

- k = number of items

- $n(k)$ = total number of responses

- Minimum $R = .80$

Errors (E) represent 1s above the diagonal and 0s below.

subject	Q1	Q2	Q3	Score
1	1	0	0	1
2	0	1	0	2
3	0	0	1	3
4	1	1	1	3

$$R = 1 - (3)/4(3) = 1 - 3/12 = 1 - .25 = .75$$

B. Likert Scaling (Summated Ratings):

- Often used to assess attitudes toward a specific topic.
- Rate agreement or disagreement with a series of statements selected to tap a specific topic.
- Typically a 5 or 7 pt scale : Agree vs. Disagree / Approve vs. Disapprove
 - each scale point is weighted: Agree = 5, Disagree = 1.
- Score = sum or average of the selected scale point weights.
- ½ the items are worded in the positive direction and ½ in the negative direction.
 - reverse score negative items before summing.
Reverse Score = $(H + L) - X$
H = high anchor L = Lowest Anchor
X= respondent's item score

Likert Scale of Graduate Student Personality Orientation Test (GSPOT).

Please rate following items with respect to how much you agree with each item

(1)	(2)	(3)	(4)	(5)
Strongly Disagree	Disagree	Neither or Both	Agree	Strongly Agree

1. I am not a compulsive neurotic. (rev.)
2. I like my imagination crushed into dust.
3. Using jargon and citing authorities is not my idea of a good time. (rev.)
4. I feel a deep need to continue the process of avoiding life.

(note: criterion validity can be established for item 1 by asking respondents of indicate their answer to the 13th decimal place.)

- Criterion for a Likert Scale

- a. Create large pool of items representing extreme high and low aspects of the dimension.
- b. Give all items to large pool of respondents (100+). For factor analyses use 5-10 Ss per item.
- c. Item Selection.

Likert originally used a Group Difference Procedure: High Group (top 25%) vs Low Group (bottom 25%). Items with largest mean differences were retained.

Today- Examine item-total correlations. Retain items that are .30 or higher.

- Factor Analysis – retain items loading on the factor .50 or greater

- can be used multidimensional scales

- d. Test internal consistency (Cronbach's Alpha) of remaining items .70 at least.

This one goes to II



C. Thurstone Scales

Provides Equal Interval Rating categories

1. Generate large number of items (200-300) representing the entire range of the dimension (highly positive to highly negative).
2. 20-30 judges "objectively" sort items using an 11 pt equally-appearing interval scale representing the degree of favorability of the items.
3. Average of judges ratings = item weighting
 - 2 items with lowest variance (high judge agreement) are select for each scale point (11 point scale)
 - the resulting scale items represent the entire continuum.
4. -Participants respond to the scale by checking items they agree with.
 - Score is the sum of the weights for the checked items.

ARE YOU READY FOR GRAD SCHOOL?

Indicate whether you agree with the following statements by checking yes or no.

Y N

1. I am rarely consumed by feelings of self doubt and loathing. (1.2)
2. I have difficulty striving for mediocrity. (2.3)
3. I hate working late into the night. (5.1)
4. I enjoy spending long hours living wretchedly. (7.6)
5. My research will make the world a better place. (9.4)
6. Originality has no place in my life. (11)

Note: In the actual presentation of the scale items the order is random, there should be 2 items per scale point (22 total) and the weightings would not be displayed.

Problems.

- Judges' ratings change over time
- Assumes that judges attitudes don't influence the placement of items.
- Tends to have lower internal reliability compared to Likert Scaling.
- Can only use unidimensional scales
- no subscales
- takes a lot of time/effort to develop scale.

4. **Semantic Differential (Osgood)**

- Can be graphical or segmented graphical
- Evaluation = good-bad, clean-dirty, beautiful-ugly
- Activity = fast-slow, active-passive, hot-cold
- Potency = strong-weak, large-small, thick-thin

(from Fishbein & Ajzen, 1975, p.76)

Evaluation is most common.

Develop multi items by selecting related evaluative pairs

7 pts scale is common, -3 to +3 , 0 = neutral

- Average scores for item across the different rating dimensions.
- Advantage is that you can skip the item selection procedures needed for Likert scales

Example: Evaluative Dimensions

Please mark the point on the scales below that best reflect how you feel about Graduate School.

Graduate School -

Clean	___, ___, ___, ___, ___, ___, ___	Dirty
Bad	___, ___, ___, ___, ___, ___, ___	Good
Beautiful	___, ___, ___, ___, ___, ___, ___	Ugly
Vomitous	___, ___, ___, ___, ___, ___, ___	Appetizing*
Dank	___, ___, ___, ___, ___, ___, ___	Shwag*

*item loadings on evaluation factor are unknown.

- Krapman's Semantic Undifferential

Graduate School-

Sucks	___ ___ ___ ___	Sucks
Blows	___ ___ ___ ___	Blows
Bitter	___ ___ ___ ___	Bitter
Hate	___ ___ ___ ___	Hate
Soul Crushing	___ ___ ___ ___	Soul Crushing
Blinding Rage	___ ___ ___ ___	Blinding Rage
Obsequious Glee	___ ___ ___ ___	Obsequious Glee

- All scale points weighted as 4.
- Score is derived by adding 6 to every response dividing by the square root of the naperian constant, hopping on one foot and punching the nearest academic in the face.
- Popularity of the scaling method has waned in todays litigious climate, despite the high internal and test-retest reliability and its well established validity in grad student populations

Recommended Readings for Scale Development

Fishbein, M., & Ajzen, I (1975). Chapter 3: Measurement Techniques. In M. Fishbein & I. Ajzen. *Beliefs, attitudes, intentions, and behaviors: An introduction to theory and research* (pp. 53 – 106).

Dawis, R. V. (1987). Scale Construction. *Journal of Counseling Psychology*, 34(4), 481-489.

V Problems with Self-Report Formats

A. Response style(set)/ Bias

1. Dispersion Bias (Dawis, 1987) = Tendency to constrain or expand the distribution of ratings

Extremity = using only endpoints, Categorical perception and Unfamiliarity with Dimensional Constructs

- fem more extreme than males
- children/Adolescents > adults
- non-college > college grad
- Af am. > cauc.

Restriction = Use only a small portion of the scale

2. Location Bias (Dawis, 1987) = responses hover at, above, or below the mean
- a. **Response Acquiescence / Deviation**
= Yea-saying / Nay-saying
- respondents are unsure how to respond
 - Social Desirability / Reactance
 - reversed items help limit
- b. **Central Tendency** = sometimes participants overuse midpoint of scale when they feel the dimension is not descriptive or unusual

3. Correlation Bias (Dawis, 1987) =
- a. **Halo effects** = ratings of one characteristic affect rating of other (often unrelated) characteristics. (like Anchoring: start high - stay high, start low - stay low)
- **Leniency/Harshness** = your evaluation of the person affects your evaluation of their characteristics.
- b. **Sequential error** = order of questions can influence responding.
- c. **Logical error** = rater coordinates responses to unrelated items

B. Judgement Heuristic Bias =

1. **Recency/Accessability Error** = recent events have disproportionate influence on judgments (especially for frequency and likelihood estimates).

The easier it is to think of an example the higher the likelihood rating people will make.

2. **Anchoring and Adjustment Error** = Scale labels can provide anchors for frequency judgment, people are quite poor at sufficiently adjusting away from anchors.

(# of African Nations in U.N. greater than or less than 64 (vs. 12) - how many nations in U.N?)

C. Social Desirability - not wanting to look bad or be uncooperative. (e.g. Crown Marlow Scale)

Two Factors identified by different Social Desirability Tests.

- Impression Management= creating a positive social image (controlled)
- Self-Deceptive Positivity= an consciously honest but overly positive presentation of the self (automatic)