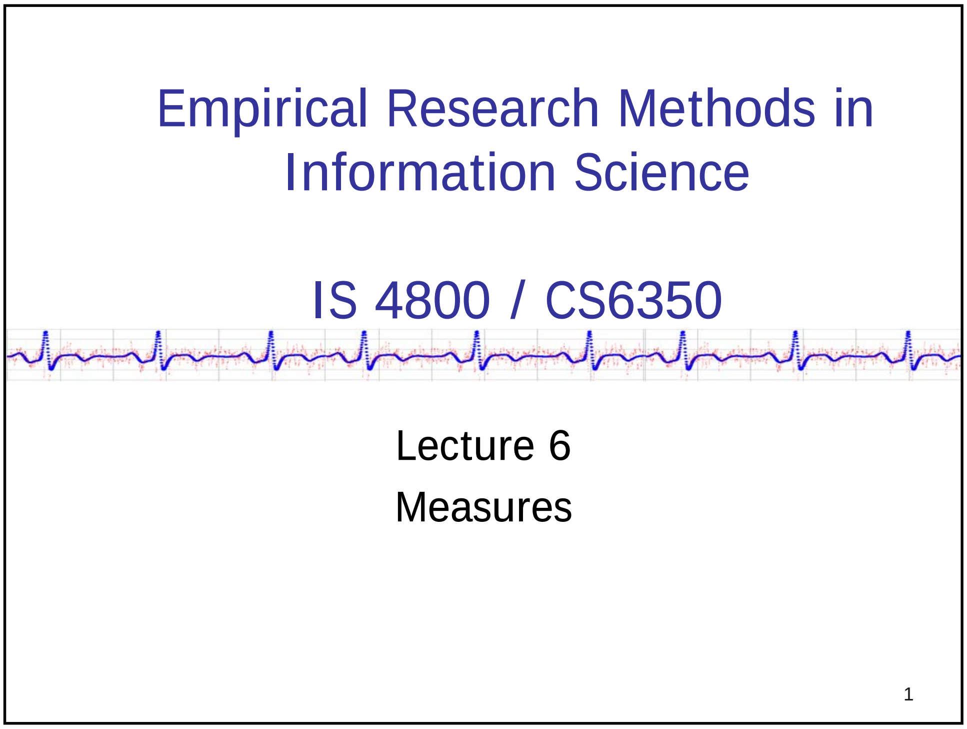
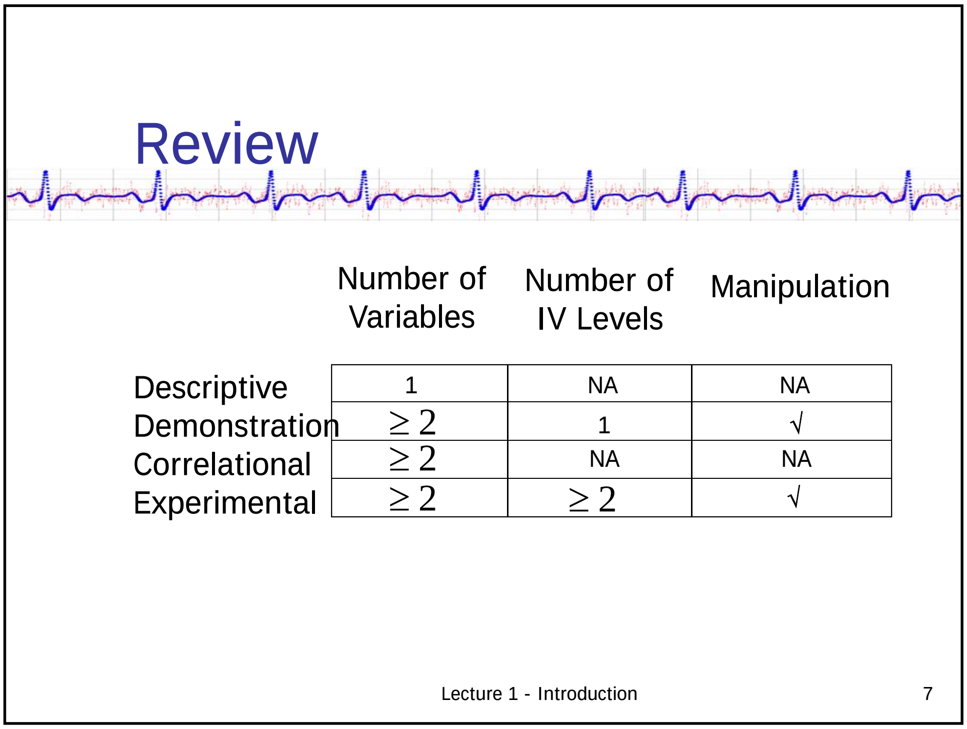


Empirical Research Methods in Information Science

IS 4800 / CS6350

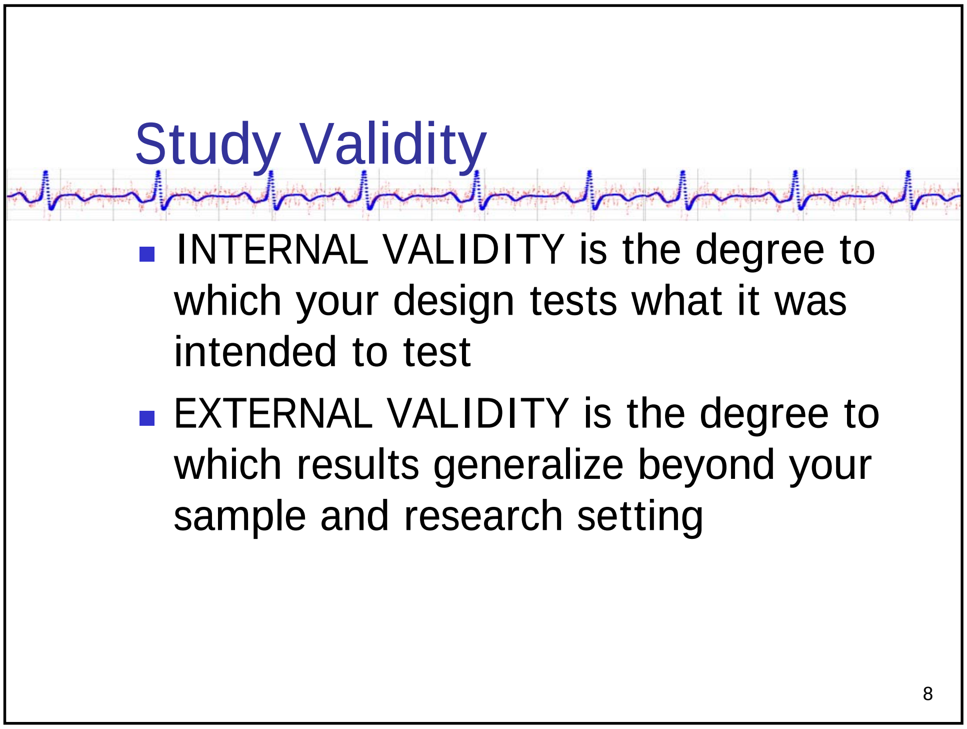
Lecture 6

Measures



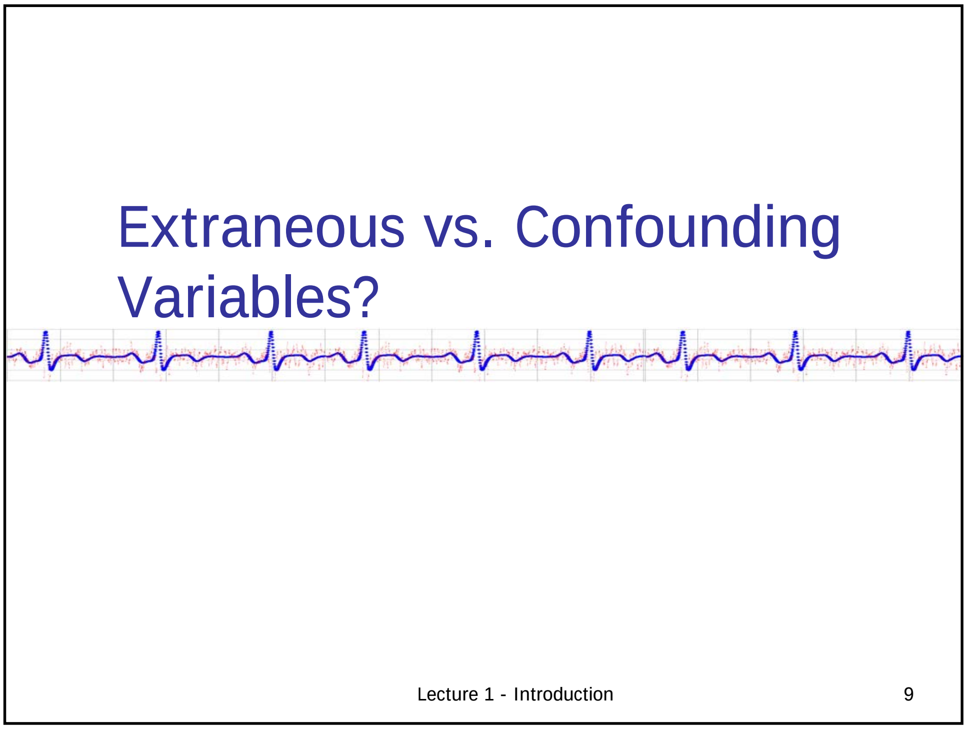
Review

	Number of Variables	Number of IV Levels	Manipulation
Descriptive	1	NA	NA
Demonstration	≥ 2	1	√
Correlational	≥ 2	NA	NA
Experimental	≥ 2	≥ 2	√



Study Validity

- INTERNAL VALIDITY is the degree to which your design tests what it was intended to test
- EXTERNAL VALIDITY is the degree to which results generalize beyond your sample and research setting



Extraneous vs. Confounding Variables?

Homework Review

Ethnography

Research Models

Making Systematic Observations

What to Measure / How to Measure it?

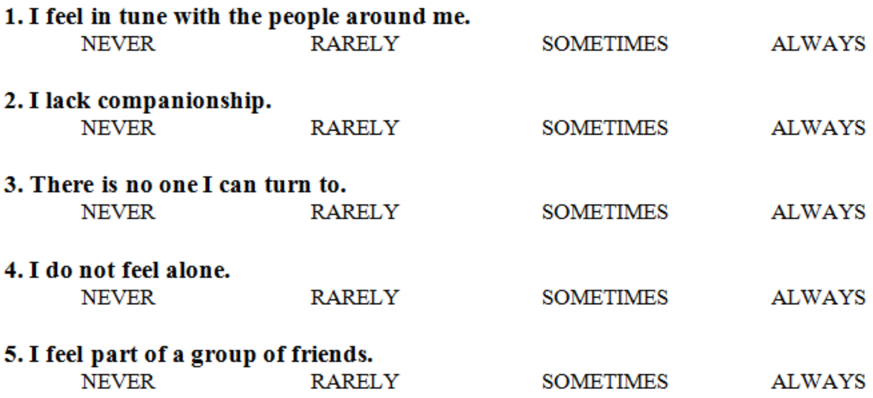
- Given the choice
  - Use a validated measure
  - Use a measure that has been used before in your field
  - Use a measure that is readily accessible or inexpensive
  - Use a measure that takes the least time and effort

What is a validated measure?

- Has reliability
- Has validity

- For psychological measures, these are collectively referred to as a measure’s “psychometrics”.

Example ‘Composite Scale Questionnaire’ UCLA Loneliness Scale (excerpt)



1. I feel in tune with the people around me.

NEVER	RARELY	SOMETIMES	ALWAYS

2. I lack companionship.

NEVER	RARELY	SOMETIMES	ALWAYS

3. There is no one I can turn to.

NEVER	RARELY	SOMETIMES	ALWAYS

4. I do not feel alone.

NEVER	RARELY	SOMETIMES	ALWAYS

5. I feel part of a group of friends.

NEVER	RARELY	SOMETIMES	ALWAYS

Measure Reliability

- A reliable measure produces similar results when repeated measurements are made under identical conditions
- Reliability can be established in several ways
  - Physical measures = precision
  - Behavioral measures = interrater reliability
  - Questionnaire measures …

Questionnaire Reliability

- *Test-retest reliability:* Administer the same test twice (or many times)
- *Parallel-forms reliability:* Alternate forms of the same test used
- *Split-half reliability:* Parallel forms are included on one test and later separated for comparison

Questionnaire Reliability

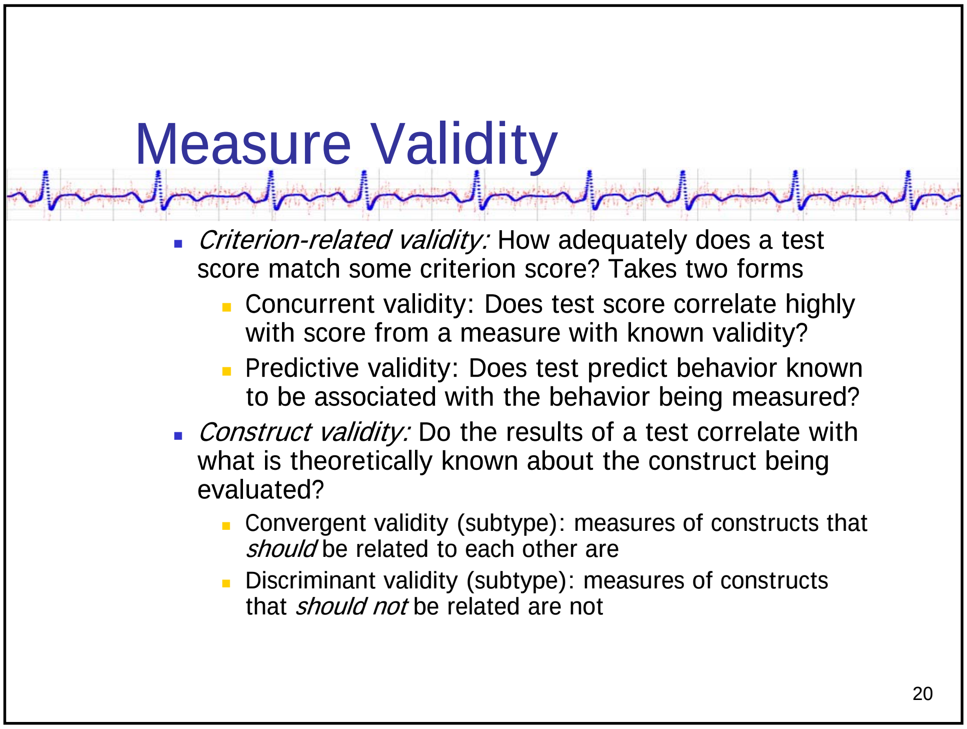
- For questionnaires using multiple questions to assess the same underlying factor, this also encompasses *internal consistency:*
  - Do all of the questions address the same underlying construct of interest?
  - That is, do scores covary?
  - A standard measure is Cronbach’s alpha
    - 0 = no correlation
    - 1 = scores always covary in the same way

Accuracy of a Measure

- *ACCURACY* of a Measure
  - Term usually applied to physical measures
  - An accurate measure produces results that agree with a known standard (i.e. is “correct”)
  - ***Precision*** is reflected in the amount of information (level of detail)
  - A measurement instrument can be inaccurate but reliable
    - The reverse cannot be true

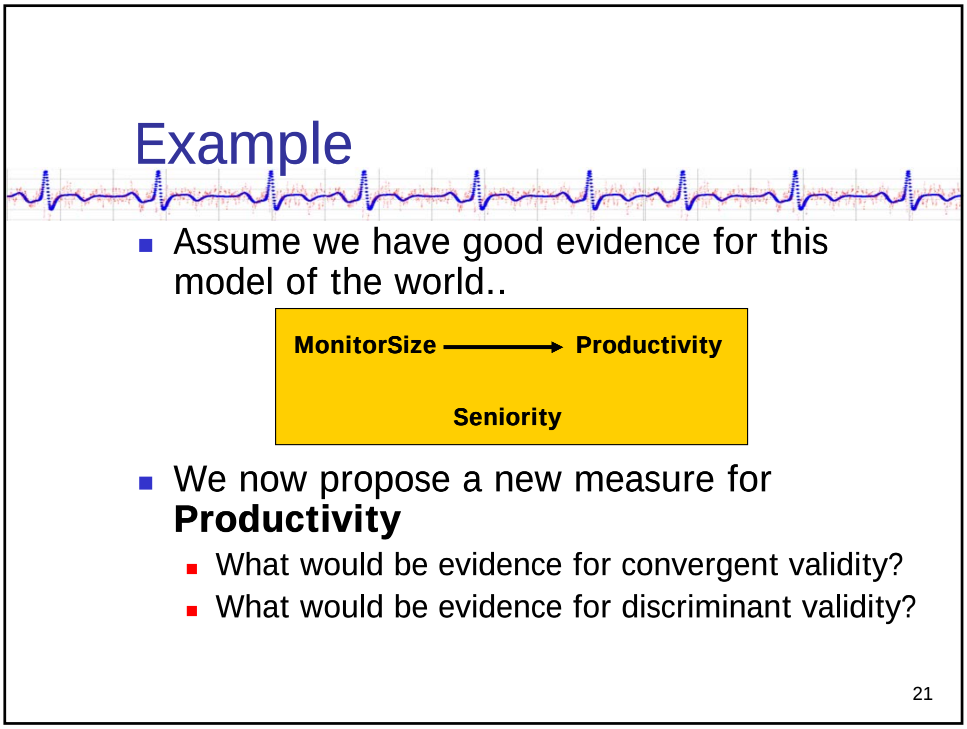
Measure Validity

- A valid measure measures what you intend it to measure
- Most important when using psychological tests (e.g., IQ test)
- Validity can be established in a variety of ways
  - *Face validity:* Assessment of adequacy of content. Least powerful method
  - *Content validity:* How adequately does a test sample behavior it is intended to measure?



Measure Validity

- *Criterion-related validity:* How adequately does a test score match some criterion score? Takes two forms
  - Concurrent validity: Does test score correlate highly with score from a measure with known validity?
  - Predictive validity: Does test predict behavior known to be associated with the behavior being measured?
- *Construct validity:* Do the results of a test correlate with what is theoretically known about the construct being evaluated?
  - Convergent validity (subtype): measures of constructs that *should* be related to each other are
  - Discriminant validity (subtype): measures of constructs that *should not* be related are not



Example

- Assume we have good evidence for this model of the world..

MonitorSize → **Productivity**

Seniority

- We now propose a new measure for **Productivity**
  - What would be evidence for convergent validity?
  - What would be evidence for discriminant validity?

Example: How good is it?

- **Diabetes Knowledge**. Diabetes knowledge will be assessed using the Diabetes Knowledge (DKN) Scales, three separate 15-item multiple choice questionnaires that measure general diabetes knowledge. Reliability for the items in the scales (Cronbach's alpha) was 0.92, indicating high internal consistency. Validity was assessed by determining that 219 participants who participated in a 1-1/2 day class on diabetes scored significantly higher posttest on the measures compared to pretest (11.27 vs. 7.61, p<.001).

Example: How good is it?

- **Fitness & Mobility** will be assessed using timed maximal walking velocity. This measure, already assessed routinely for all GAP patients, involves having subjects walk along an 11-meter, straight, flat walkway as fast as possible. Each subject will have three trials, with 30-second intervening rest periods. The time taken to walk from the 3-m to the 8-m mark on the walkway is determined, and the highest velocity among the trials is used. Maximal walking velocity was found to be significantly correlated with both peak knee-extension torque (r>0.90, p<.05) and VO2max (r>0.80, p<.05).

Example: How good is it?

- **Loneliness** will be assessed using the UCLA Loneliness Scale. This measure is highly reliable, both in terms of internal consistency (alpha ranging from .89 to .94) and test-retest reliability over a 1-year period (r = .73). Convergent validity for the scale was indicated by significant correlations with other measures of loneliness. Construct validity was supported by significant relations with measures of the adequacy of the individual's interpersonal relationships, and by correlations between loneliness and measures of health and well-being

Example: How good is it?

- **Exercise Self Efficacy.** The five-item Self Efficacy Scale for exercise assesses perceived confidence to perform exercise across a wide variety of challenging situations. Recently, a new measure was developed addressing the multidimensionality of the self-efficacy construct. The short form (α = .82) of this measure includes six items, answered on a five point Likert response format and assesses negative affect, excuse making, exercising alone, equipment access, resistance from others and weather.

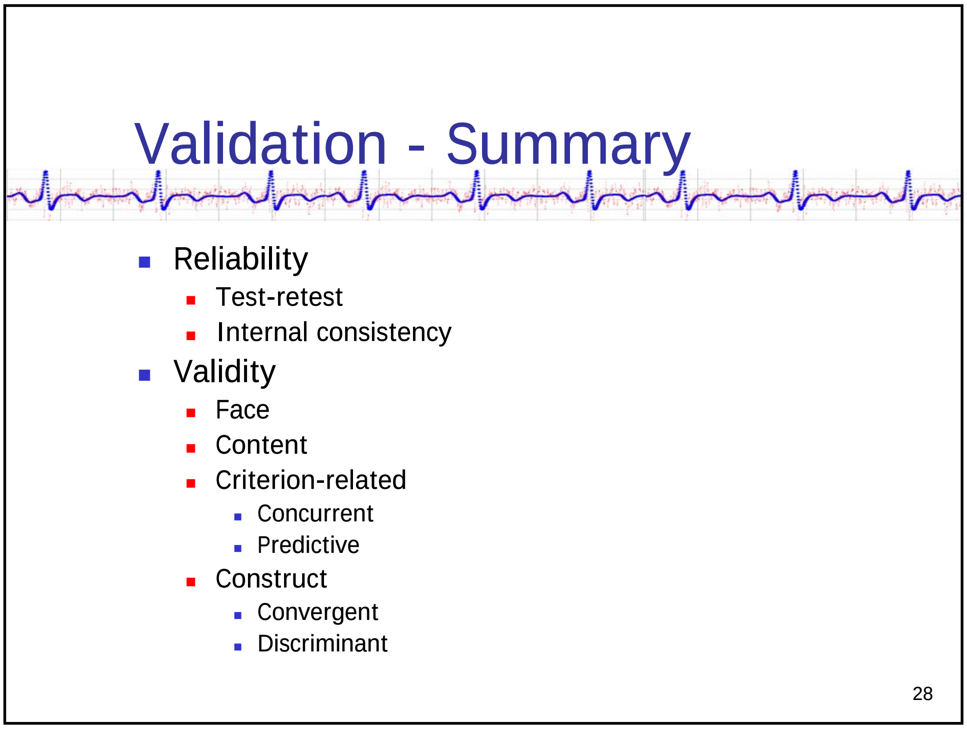
Example: How good is it?

- **Patient Activation.** Patient activation will be assessed using the Patient Activation Measure (PAM). This 22-item self-report questionnaire assesses: a) beliefs about the importance of the patient role; b) confidence and knowledge necessary to take action; c) actions actually taken; and d) ability to stay the course when under stress. In an assessment involving a national sample of 1,515 individuals aged 45 and over, the instrument was shown to have high reliability and construct validity: those with higher activation reported significantly better health as assessed by the SF-8 (r=.38, p<.001) and have significantly lower rates of doctor office visits, emergency room visits, and hospital nights (r=-.07, p<.01).

Developing a New Measure

- Say you decide you need a new survey measure, “attitude towards large computer monitors” (ATLCM)
  - I like big monitors.
  - Big monitors make me nervous.
  - I prefer small monitors, even if they cost more.
  - *7-pt Likert scales*

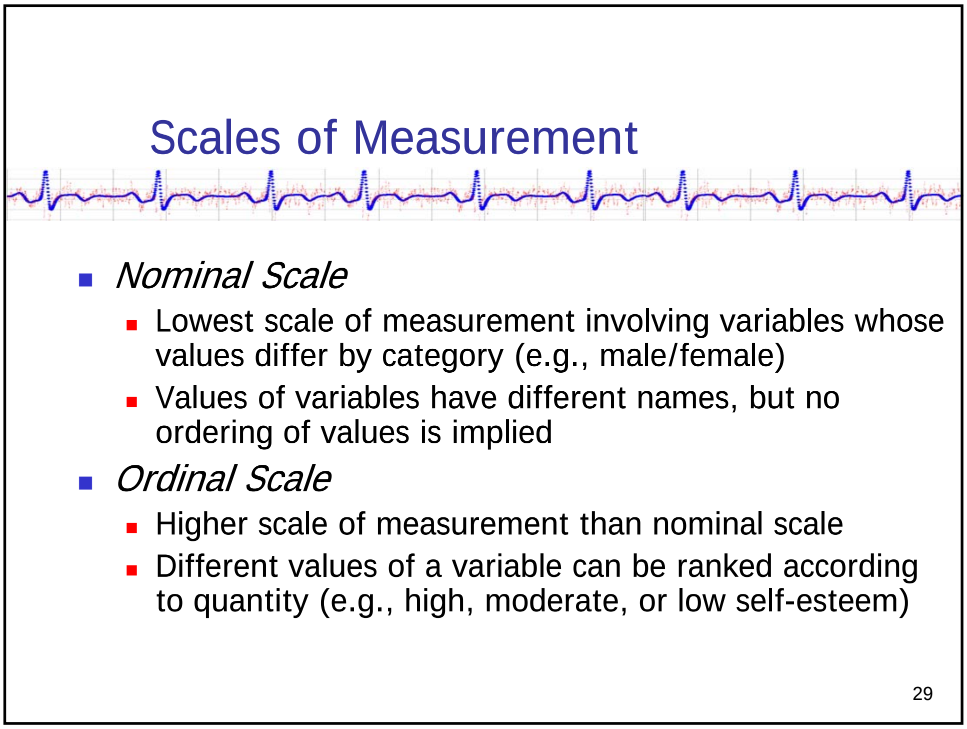
- How would you validate this measure?



Validation - Summary

- Reliability
  - Test-retest
  - Internal consistency
- Validity
  - Face
  - Content
  - Criterion-related
    - Concurrent
    - Predictive
  - Construct
    - Convergent
    - Discriminant

28



Scales of Measurement

- *Nominal Scale*
  - Lowest scale of measurement involving variables whose values differ by category (e.g., male/female)
  - Values of variables have different names, but no ordering of values is implied
- *Ordinal Scale*
  - Higher scale of measurement than nominal scale
  - Different values of a variable can be ranked according to quantity (e.g., high, moderate, or low self-esteem)

29

Scales of Measurement

- *Interval Scale*
  - Scale of measurement on which the spacing between values is known (e.g., rating a book on a scale ranging from 0 to 10)
  - No true zero point
- *Ratio Scale*
  - Similar to interval scale, but with a true zero point (e.g., number of lever presses)

What kind is it?

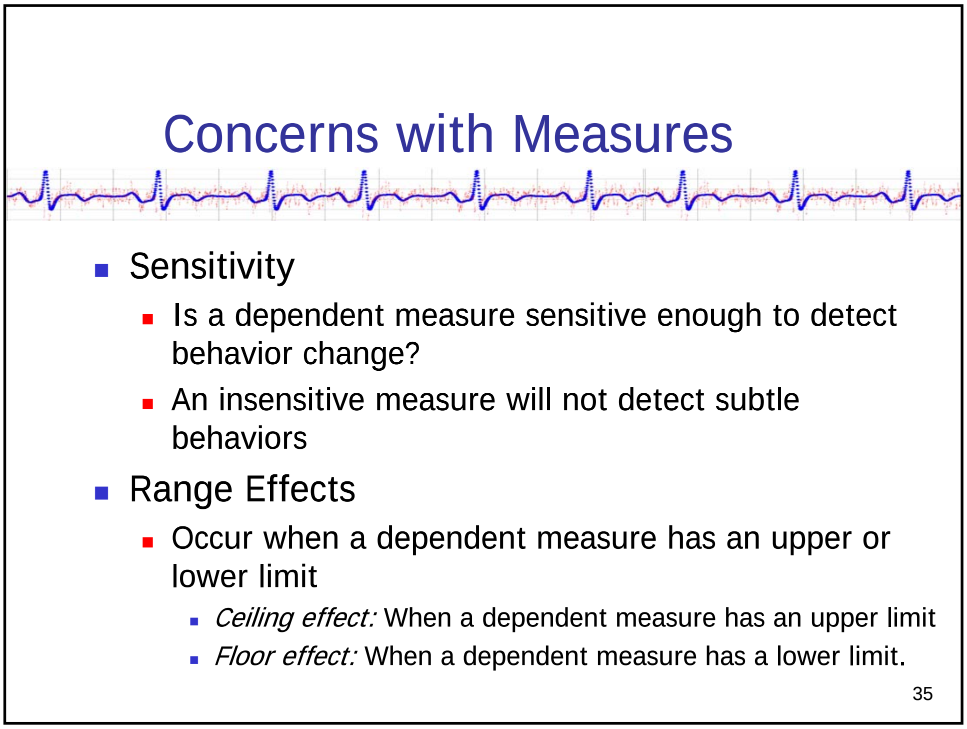
- Age
- Gender
- Job Category (Engineer, Manager…)
- Weight
- School Year (Freshman…)
- Temperature (Celsius)
- Olympic medal (Gold, Silver, Bronze)
- Monitor Size
- Weather (Rain, Snow, …)
- Salary
- Productivity (wpd)
- Owns Pet (or not)

Practically speaking

- You will decide on statistical tests depending on whether your measures are
  - Nominal or Ordinal, or
  - Numeric (Interval, Ratio)

Factors Affecting Your Choice of a Scale of Measurement

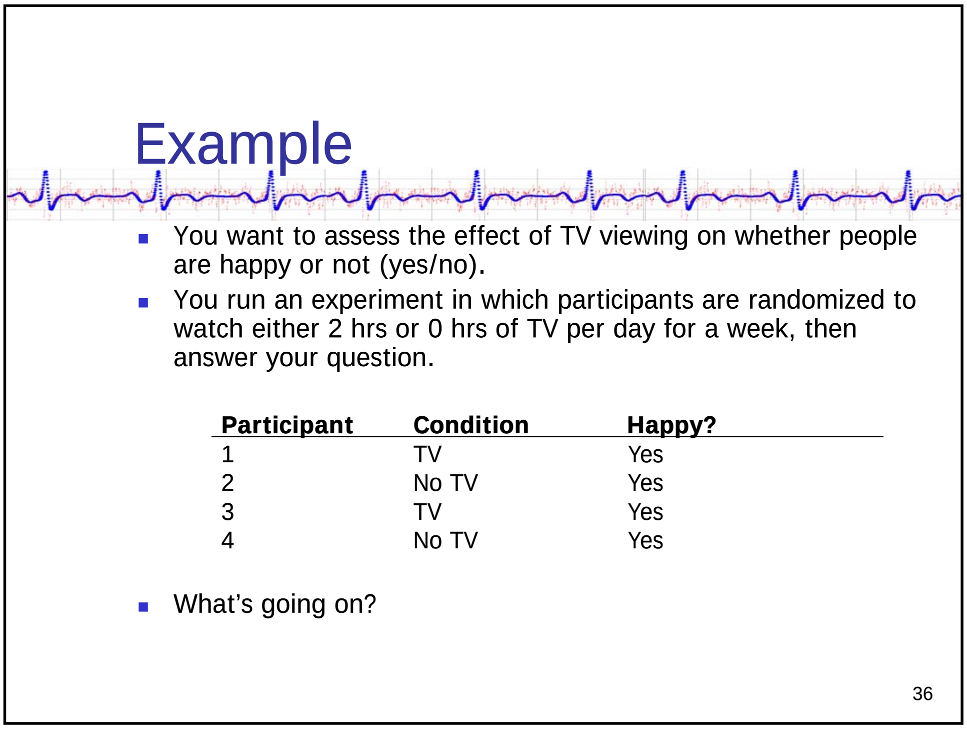
- Information Yielded
  - A nominal scale yields the least information.
  - An ordinal scale adds some crude information.
  - Interval and ratio scales yield the most information.
- Statistical Tests Available
  - The statistical tests available for nominal and ordinal data (nonparametric) are less powerful than those available for interval and ratio data (parametric)
  - Use the scale that allows you to use the most powerful statistical test



Concerns with Measures

- Sensitivity
  - Is a dependent measure sensitive enough to detect behavior change?
  - An insensitive measure will not detect subtle behaviors
- Range Effects
  - Occur when a dependent measure has an upper or lower limit
    - *Ceiling effect:* When a dependent measure has an upper limit
    - *Floor effect:* When a dependent measure has a lower limit.

35



Example

- You want to assess the effect of TV viewing on whether people are happy or not (yes/no).
- You run an experiment in which participants are randomized to watch either 2 hrs or 0 hrs of TV per day for a week, then answer your question.

Participant	Condition	Happy?
1	TV	Yes
2	No TV	Yes
3	TV	Yes
4	No TV	Yes

- What’s going on?

36

Example

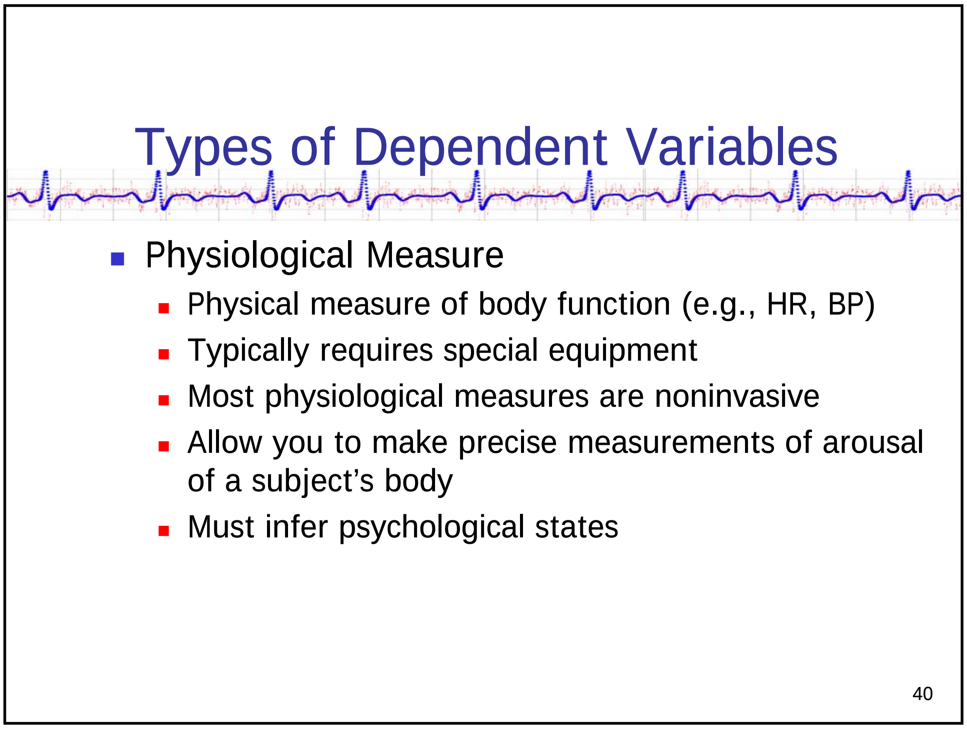
- You want to assess the effect of TV viewing on positive affect, measured on a 1-7 scale (PANAS).
- You run an experiment in which participants are randomized to watch either 2 hrs or 0 hrs of TV per day for a week, then fill out the PANAS.

Participant	Condition	PANAS
1	TV	7.0
2	No TV	6.7
3	TV	6.9
4	No TV	7.0

- What’s going on?

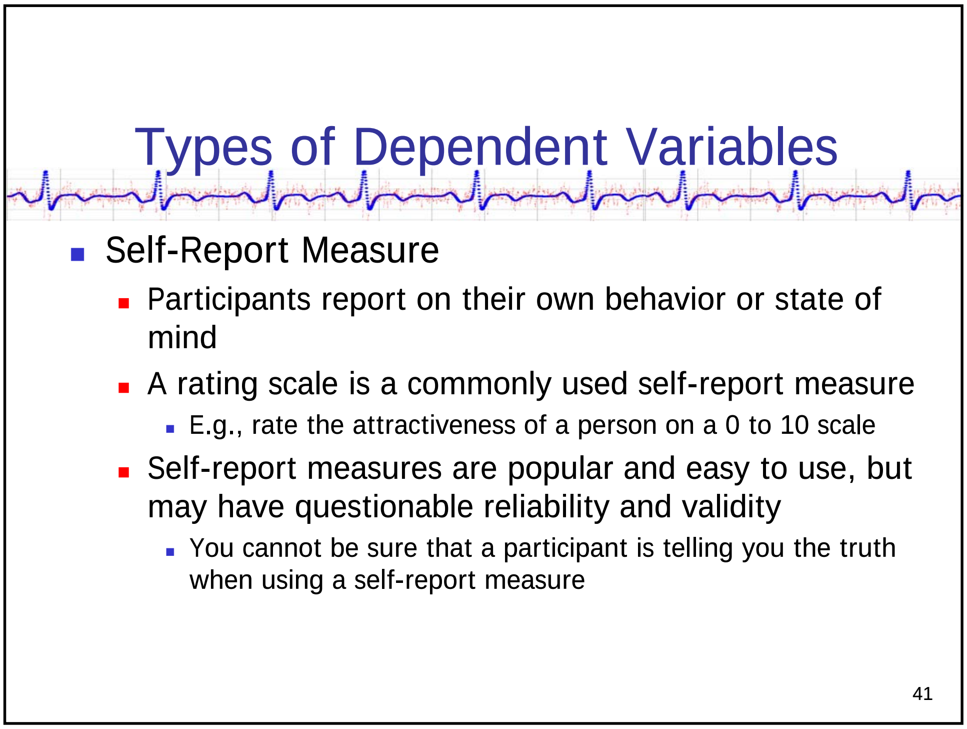
Types of Dependent Variables

- Behavioral Measure
  - Record actual behavior of subjects
  - Many, many types
    - *Frequency:* Count of the number of behaviors that occur
    - *Duration:* The amount of time it takes for a behavior to occur
    - *Number of errors:* The number of incorrect responses made
    - Subjective judgments
  - More on this next week



Types of Dependent Variables

- Physiological Measure
  - Physical measure of body function (e.g., HR, BP)
  - Typically requires special equipment
  - Most physiological measures are noninvasive
  - Allow you to make precise measurements of arousal of a subject’s body
  - Must infer psychological states



Types of Dependent Variables

- Self-Report Measure
  - Participants report on their own behavior or state of mind
  - A rating scale is a commonly used self-report measure
    - E.g., rate the attractiveness of a person on a 0 to 10 scale
  - Self-report measures are popular and easy to use, but may have questionable reliability and validity
    - You cannot be sure that a participant is telling you the truth when using a self-report measure

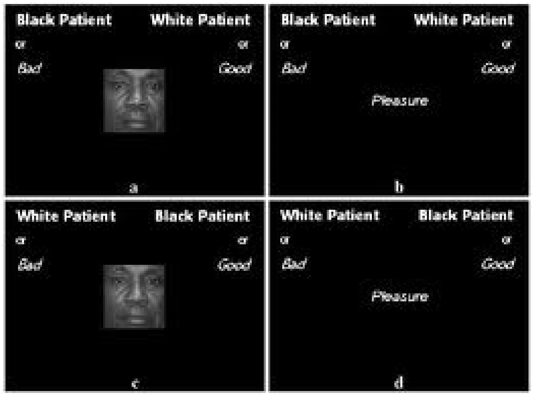
Types of Dependent Variables

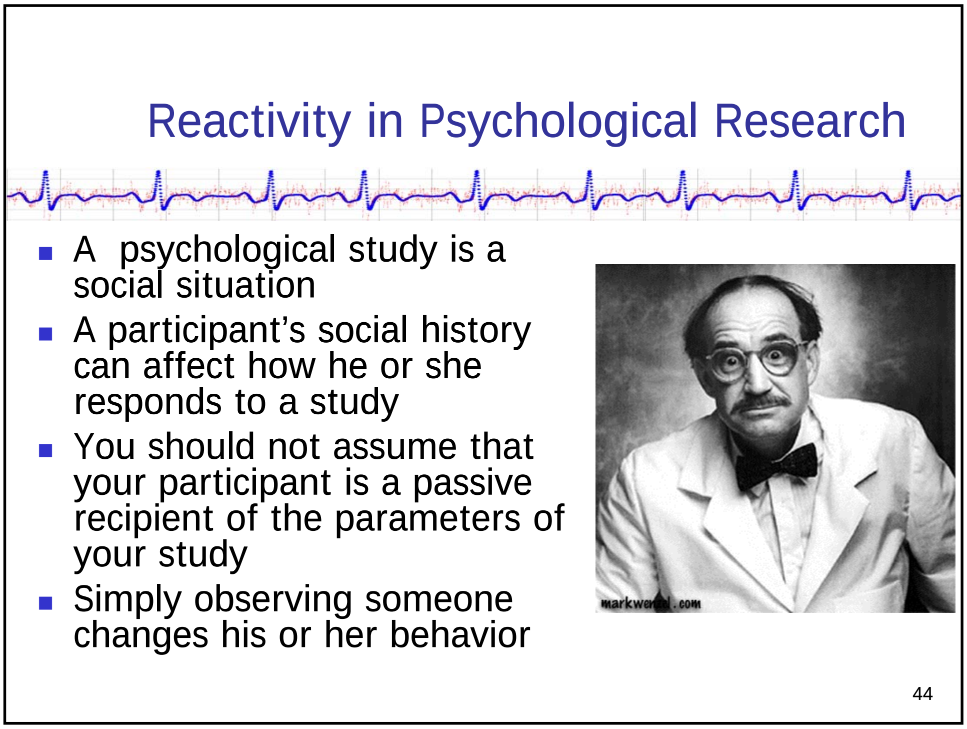
- Physical measures
  - Temperature, pressure, etc.
- System measures
  - Profiling (%use, %CPU, etc.)
  - Runtime (clock or CPU)
  - MTBF
  - etc. etc.

Implicit Measures

subject is unconscious of measurement

- Uses rapid, unconscious categorization task to tease out biases.
- Assume quicker reaction times are associated with stronger concept associations.

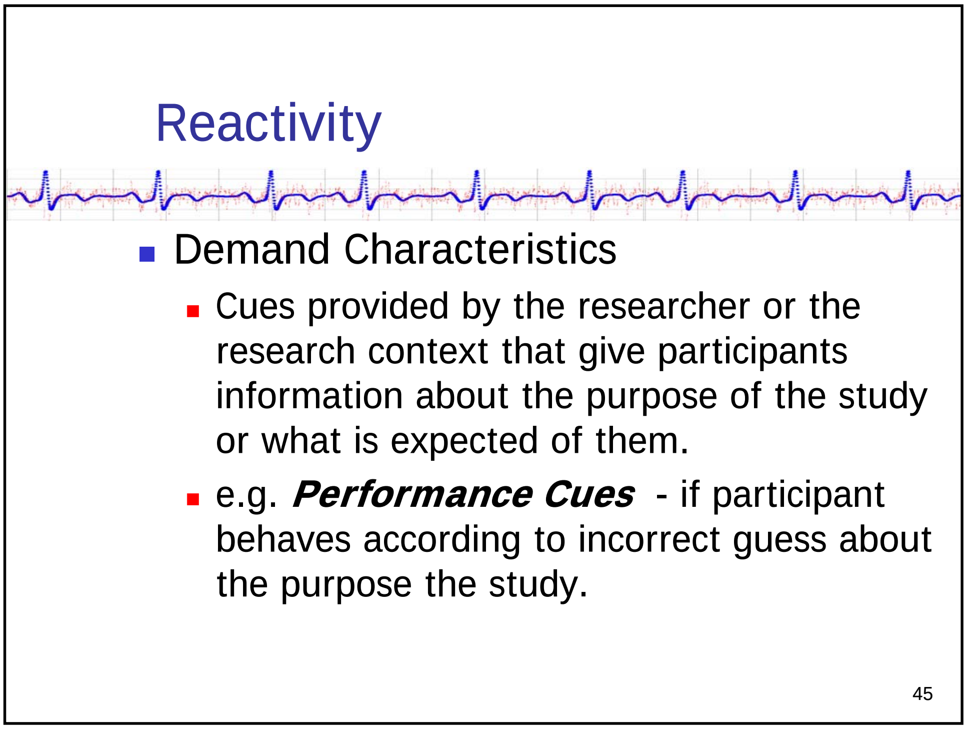




Reactivity in Psychological Research

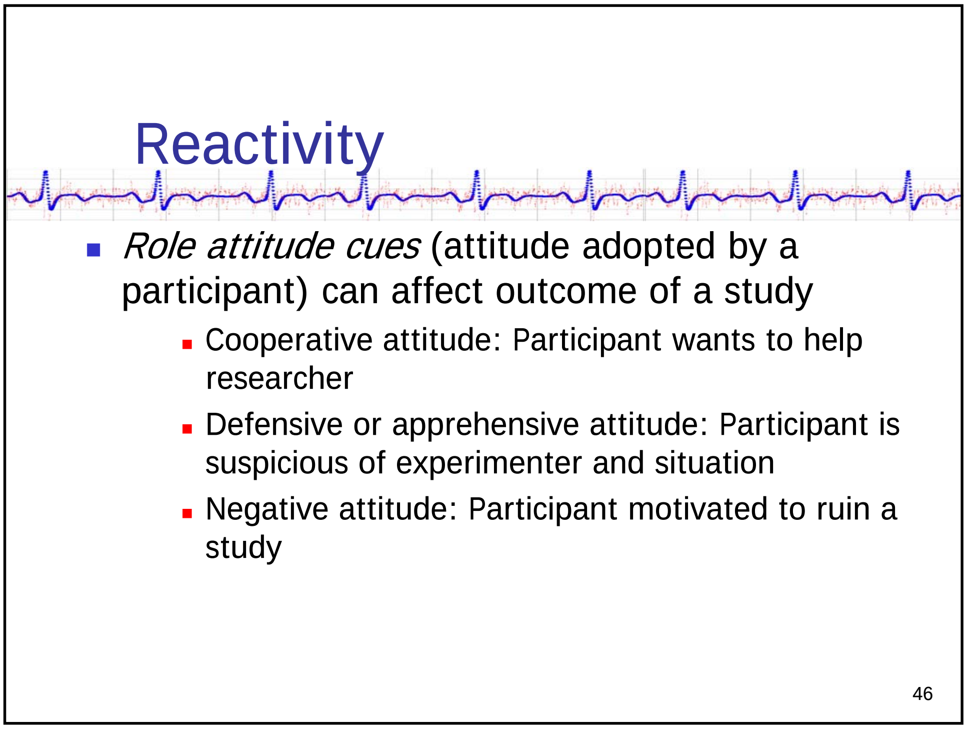
- A psychological study is a social situation
- A participant’s social history can affect how he or she responds to a study
- You should not assume that your participant is a passive recipient of the parameters of your study
- Simply observing someone changes his or her behavior

markwenzel.com



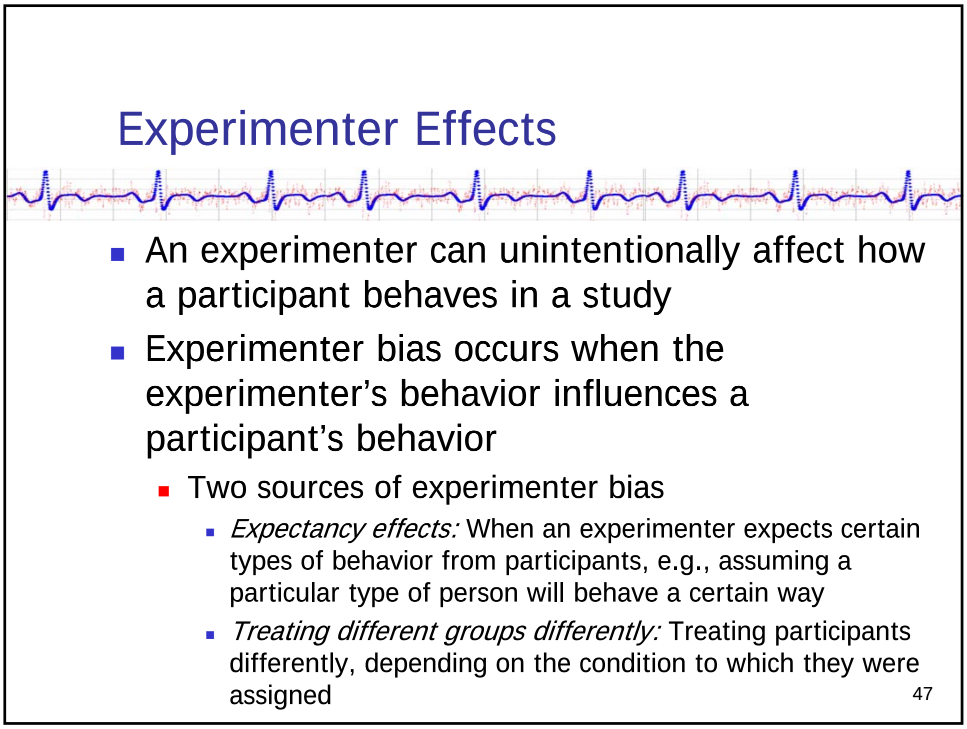
Reactivity

- Demand Characteristics
  - Cues provided by the researcher or the research context that give participants information about the purpose of the study or what is expected of them.
  - e.g. ***Performance Cues*** - if participant behaves according to incorrect guess about the purpose the study.



Reactivity

- *Role attitude cues* (attitude adopted by a participant) can affect outcome of a study
  - Cooperative attitude: Participant wants to help researcher
  - Defensive or apprehensive attitude: Participant is suspicious of experimenter and situation
  - Negative attitude: Participant motivated to ruin a study



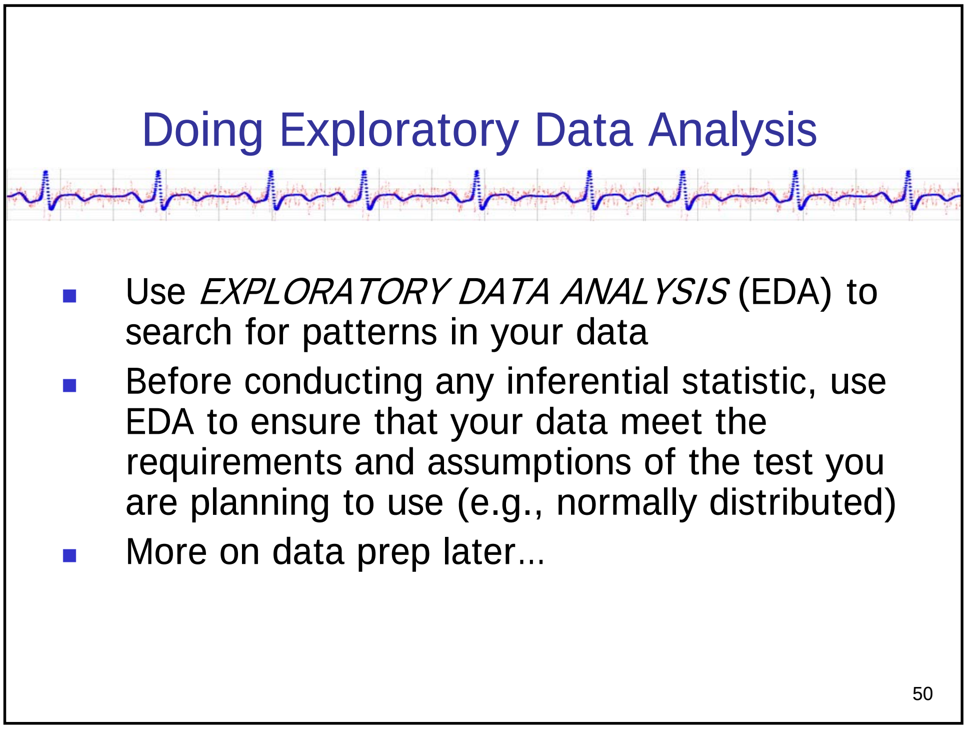
Experimenter Effects

- An experimenter can unintentionally affect how a participant behaves in a study
- Experimenter bias occurs when the experimenter's behavior influences a participant's behavior
  - Two sources of experimenter bias
    - *Expectancy effects:* When an experimenter expects certain types of behavior from participants, e.g., assuming a particular type of person will behave a certain way
    - *Treating different groups differently:* Treating participants differently, depending on the condition to which they were assigned

- Experimenter bias affects internal and external validity
- Steps must be taken to reduce experimenter bias
  - Use a *blind technique* where the experimenter does not know the condition to which a participant has been assigned
  - Use a *double-blind technique* where neither the experimenter nor participant knows the condition to which a participant has been assigned
  - Automate the experiment

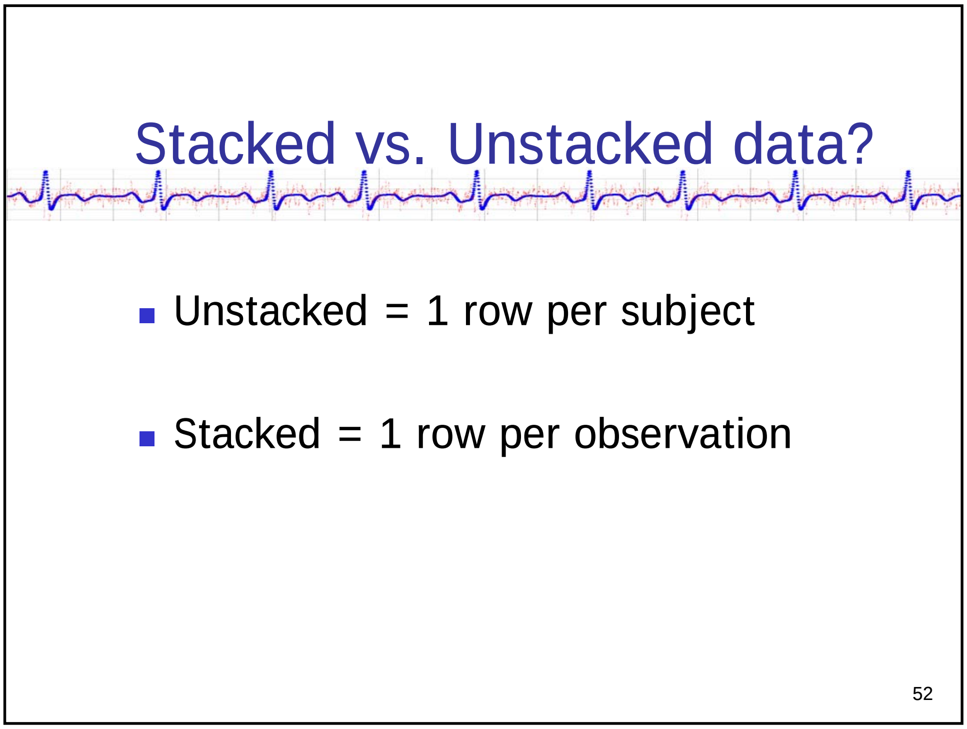
Chapter 13

Describing Data



Doing Exploratory Data Analysis

- Use *EXPLORATORY DATA ANALYSIS* (EDA) to search for patterns in your data
- Before conducting any inferential statistic, use EDA to ensure that your data meet the requirements and assumptions of the test you are planning to use (e.g., normally distributed)
- More on data prep later…



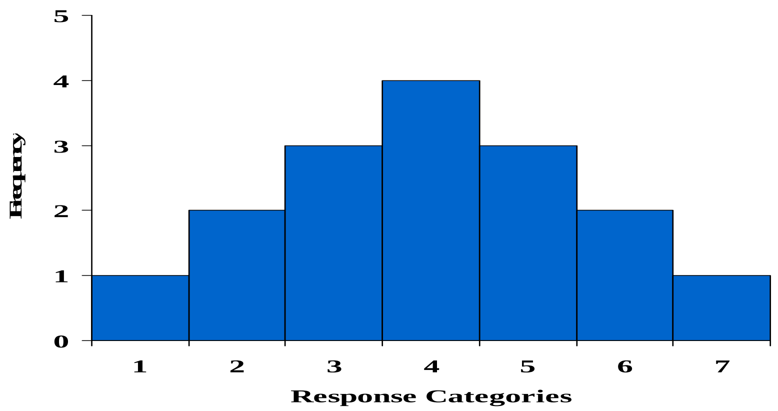
Stacked vs. Unstacked data?

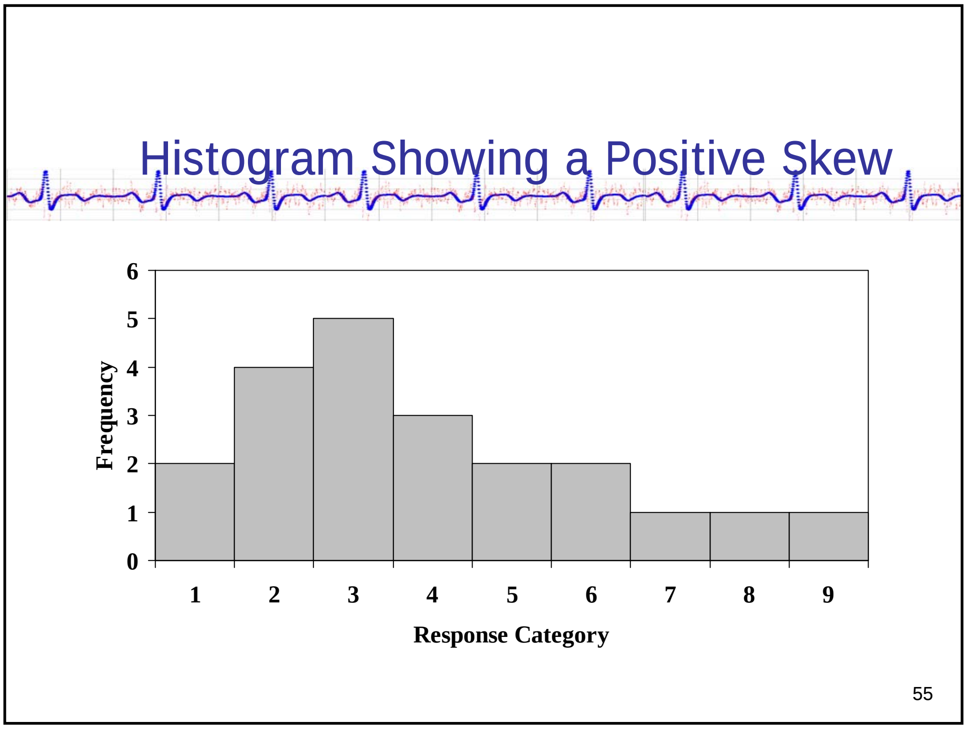
- Unstacked = 1 row per subject
- Stacked = 1 row per observation

The Frequency Distribution

- Represents a set of mutually exclusive categories into which actual values are classified
- Can take the form of a table or a graph
- Graphically, a frequency distribution is shown on a *histogram*
  - A bar graph on which the bars touch
  - The y-axis represents a frequency count of the number of observations falling into a category
  - Categories represented on the x-axis

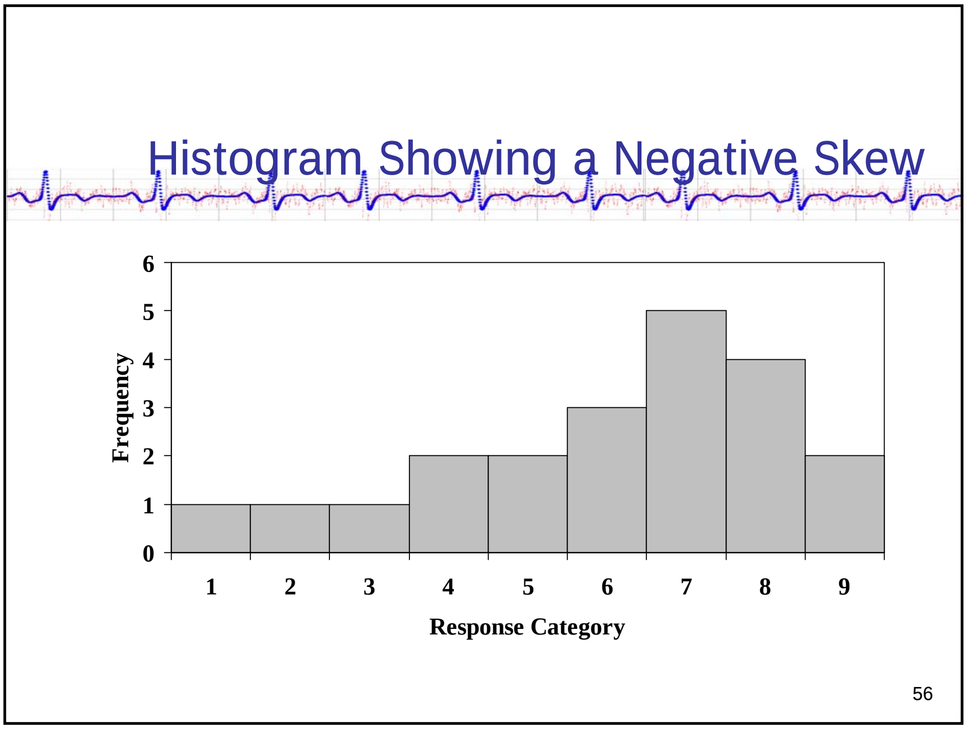
Histogram Showing a Normal Distribution





Histogram Showing a Positive Skew

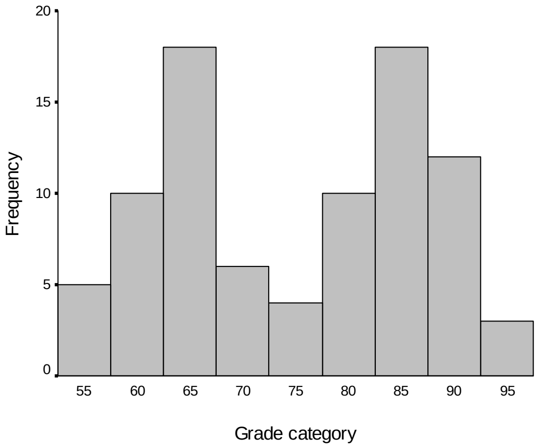
55



Histogram Showing a Negative Skew

56

A Bimodal Distribution



Measures of Center

- Mean
- Median
- Mode

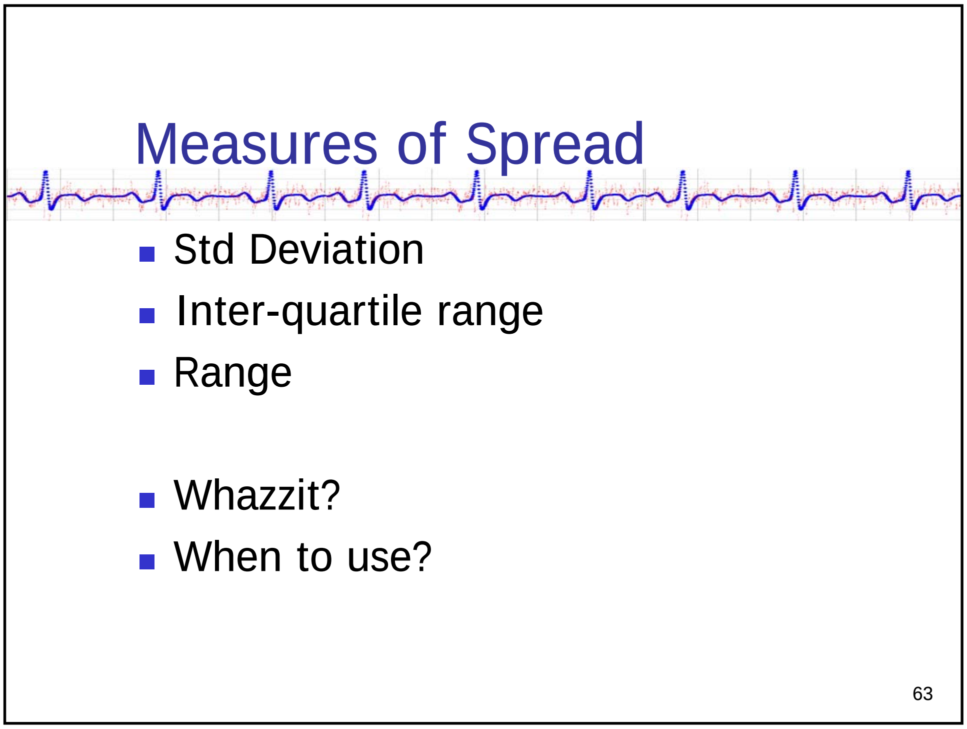
- Whazzit?
- When to use?

Measures of Center: Applications

- *Mode*
  - Used if data are measured along a nominal scale
- *Median*
  - Used if data are measured along an ordinal scale
  - Used if interval data do not meet requirements for using the mean (skewed but unimodal), or if significant outliers

Measures of Center: Applications

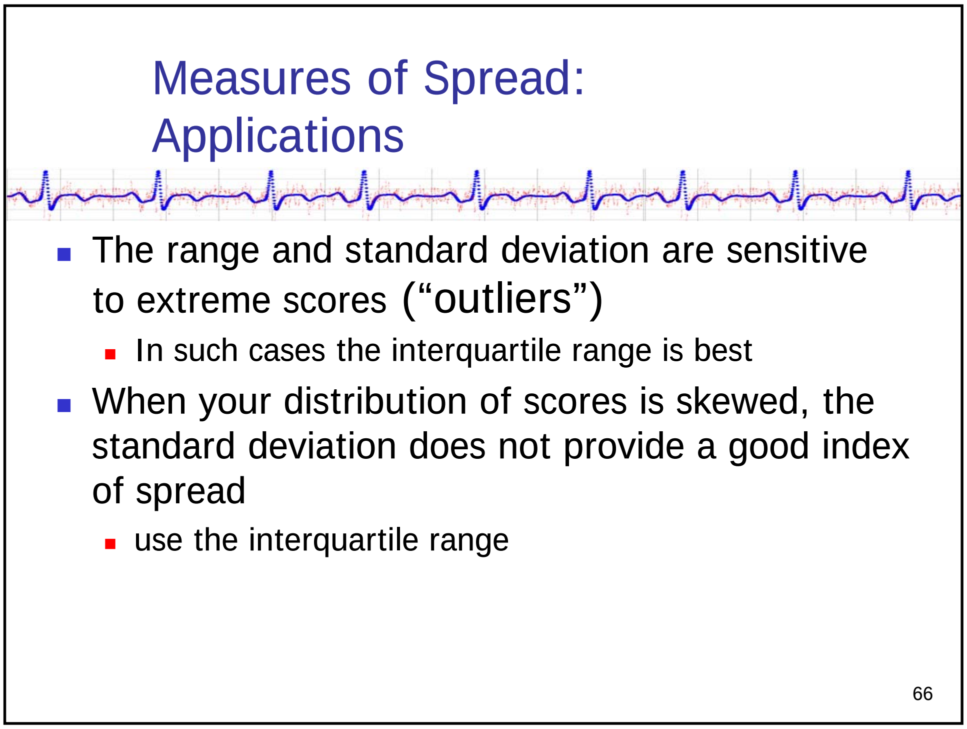
- *Mean*
  - Used if data are measured along an interval or ratio scale
  - Most sensitive measure of center
  - Used if scores are normally distributed



Measures of Spread

- Std Deviation
- Inter-quartile range
- Range

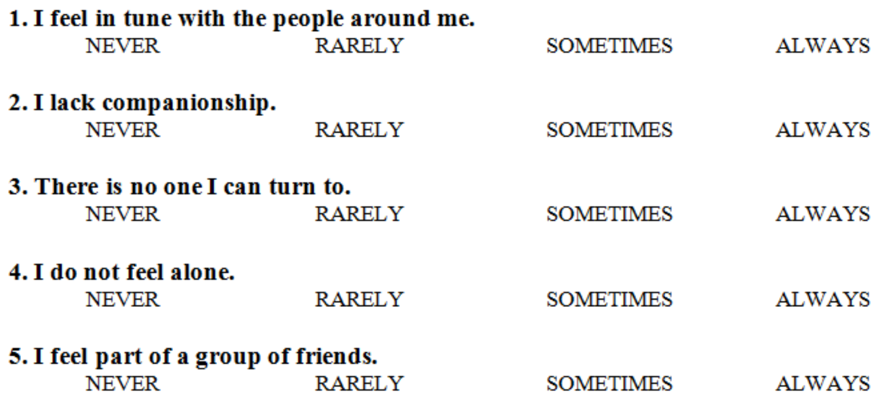
- Whazzit?
- When to use?



Measures of Spread: Applications

- The range and standard deviation are sensitive to extreme scores ("outliers")
  - In such cases the interquartile range is best
- When your distribution of scores is skewed, the standard deviation does not provide a good index of spread
  - use the interquartile range

How to analyze questionnaire data: Example ‘Composite Scale Questionnaire’



1. I feel in tune with the people around me.

NEVER RARELY SOMETIMES ALWAYS

2. I lack companionship.

NEVER RARELY SOMETIMES ALWAYS

3. There is no one I can turn to.

NEVER RARELY SOMETIMES ALWAYS

4. I do not feel alone.

NEVER RARELY SOMETIMES ALWAYS

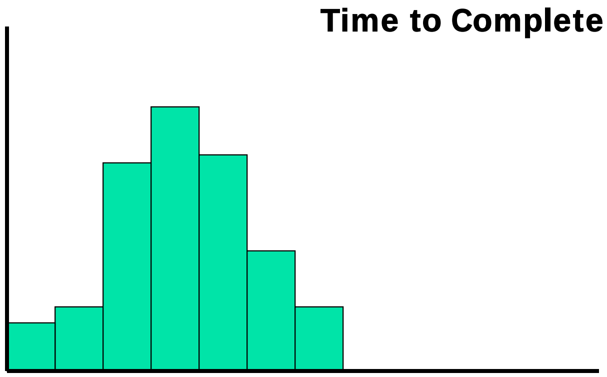
5. I feel part of a group of friends.

NEVER RARELY SOMETIMES ALWAYS

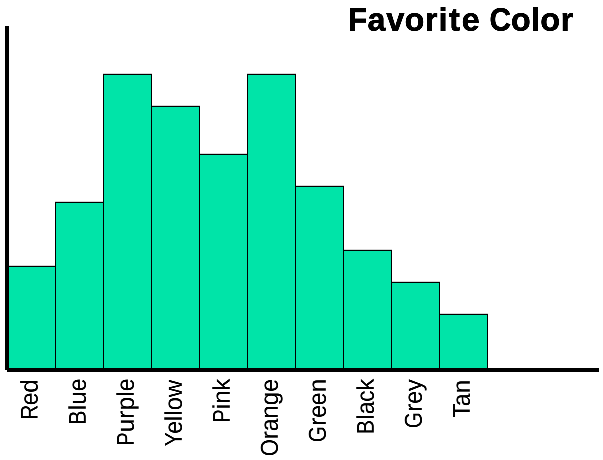
50-year controversy over analysis of scale measures

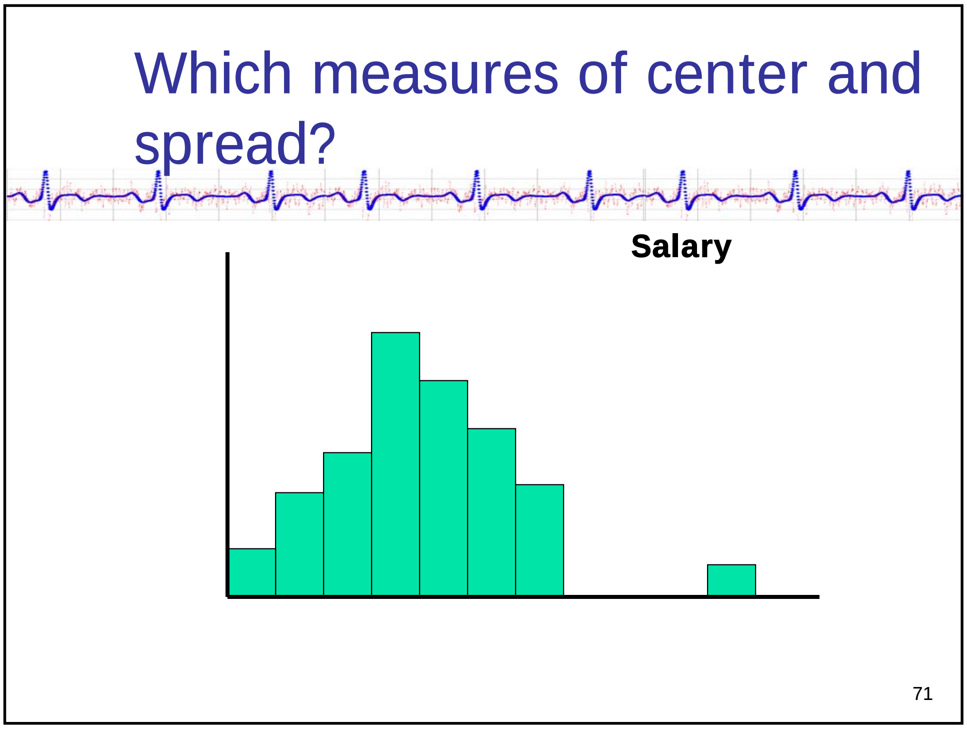
- Historically, have been treated as interval if they appear normal (i.e., with mean, stdev, and t-test)
- Some statisticians say NEVER. They are ordinal measures – must use median, no meaningful range measures, and non-parametric inferential statistics (e.g., Mann-Whitney)
- But, differentiate between scale items and composite scale measures, in which responses to multiple items are combined to yield a single numeric score.

Which measures of center and spread?

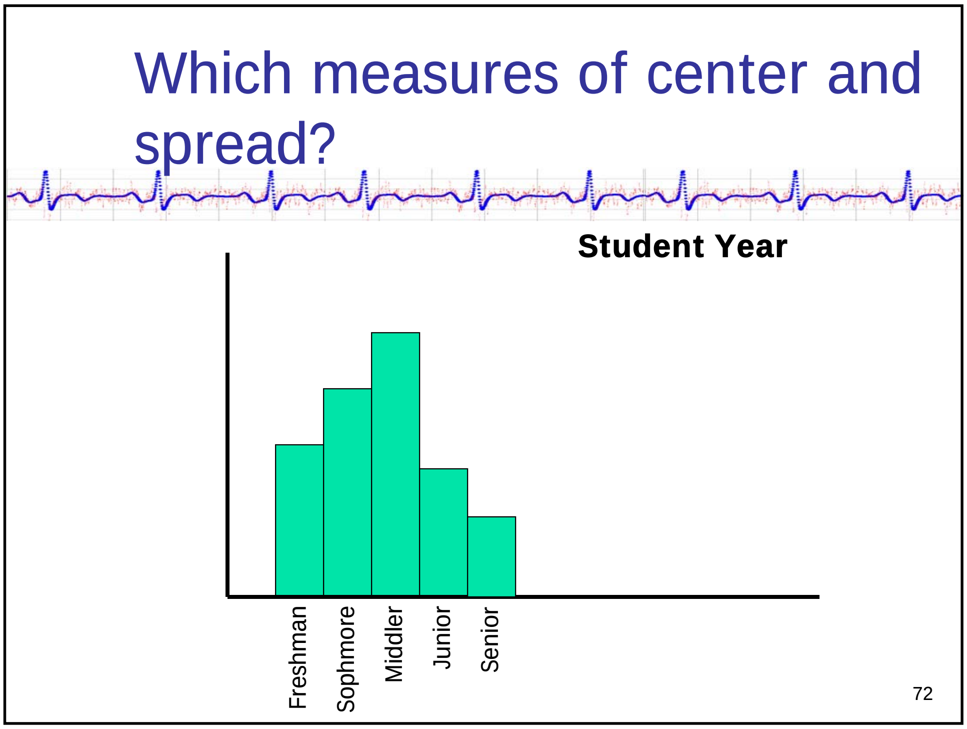


Which measures of center and spread?

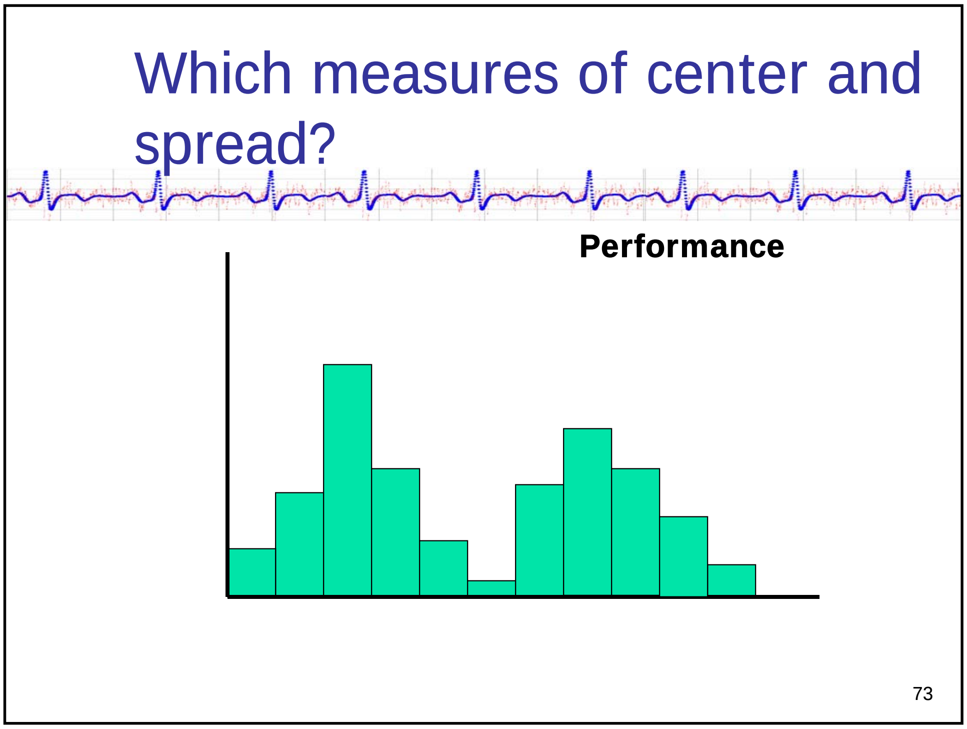




Which measures of center and spread?

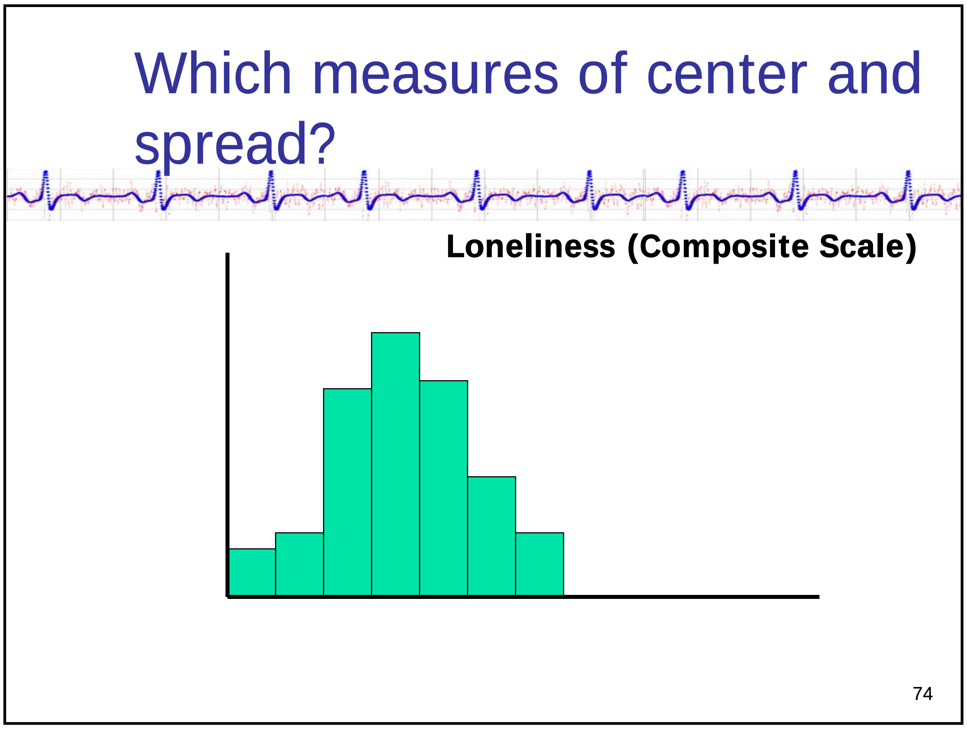


Which measures of center and spread?



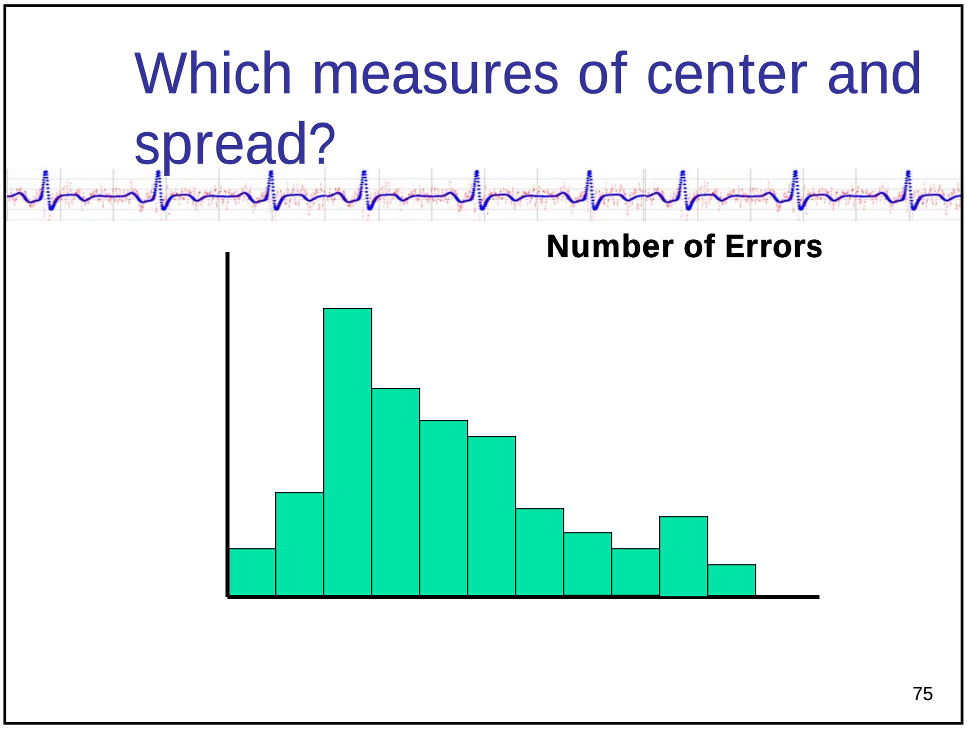
Which measures of center and spread?

Performance

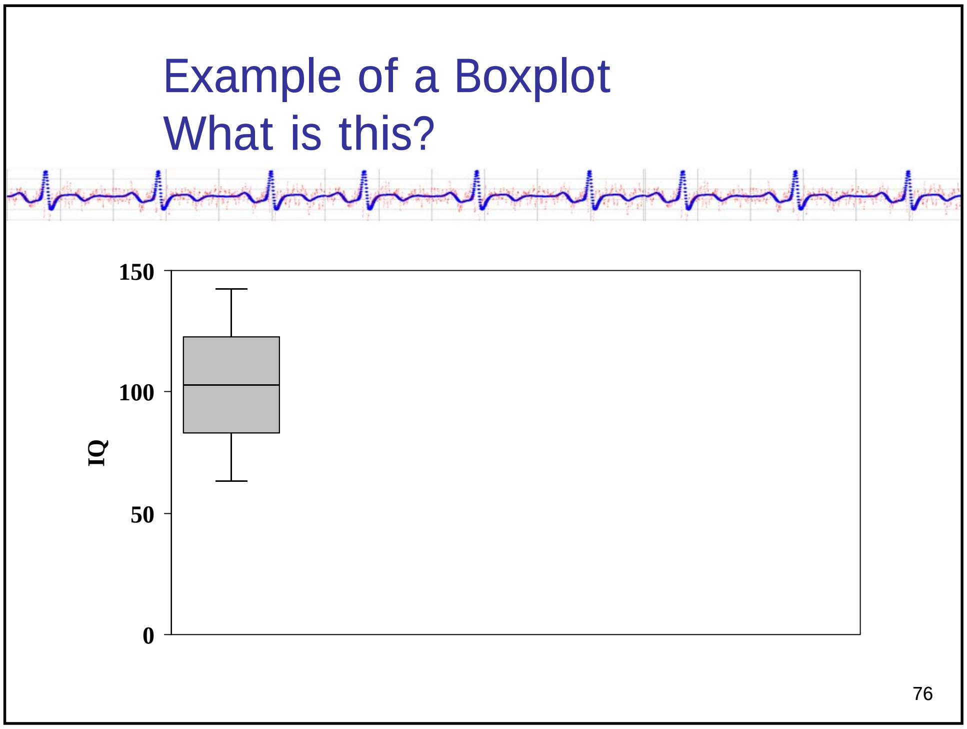


Which measures of center and spread?

Loneliness (Composite Scale)



Which measures of center and spread?



Example of a Boxplot
What is this?

Aron & Aron Nomenclature

$$M = \frac{\sum X}{N}$$

$$SS = \sum (X - M)^2$$

$$SD^2 = \frac{SS}{N}$$

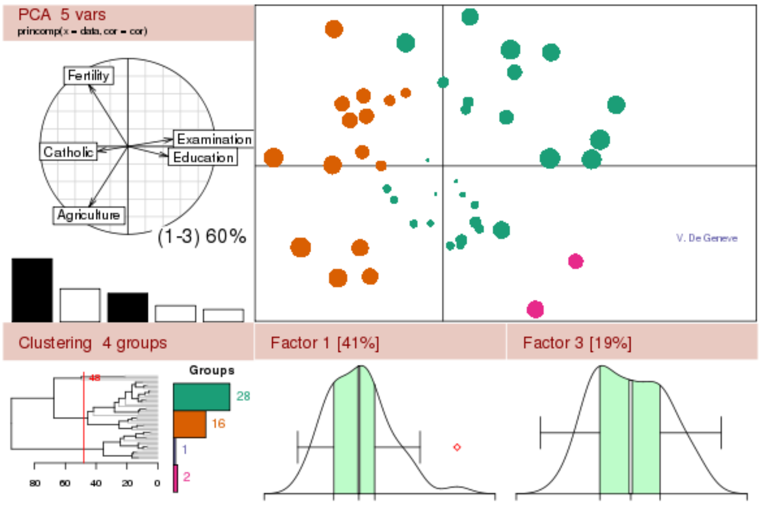
Z-scores

- Measures that have been normalized to make comparisons easier.

$$Z = \frac{X - M}{SD}$$

- Z-scores descriptives
  - Mean?
  - SD?
  - Variance?

The R Project for Statistical Computing Descriptive Statistics



Basics

- Basic descriptives

```
summary(eruptions)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.600   2.163   4.000   3.488   4.454   5.100
```

- Histogram

```
hist(eruptions)
```

More Basics

- mean(data)
- median(data)

- var(data)
- sd(data)
- IQR(data)   #inter-quartile range

Mode

- Given values in vector ‘x’

```
temp <- table(as.vector(x))
names(temp)[temp == max(temp)]
```

#this returns the names of the values that have the highest frequency in x.

Frequency Tables

- table(data) //freq counts
- prop.table(table(data)) //relative frequencies

Boxplot

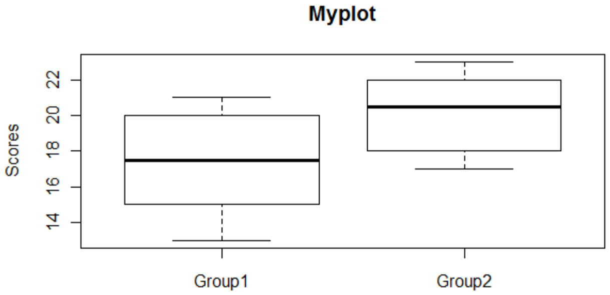
- Boxplot

  ```
  boxplot(data)
  ```

- boxplot(values~ factor) //by group

Boxplots per Group

```
> boxplot(x, y, ylab="Scores ", +
    names=c(“Group1",“Group2"),   +
```



Contingency Tables

- Co-occurrence of values for 2 variables.
- table(val1,val2)   #freq counts

Homework

- Read Nielsen Ch6 + 2 papers
- Do Homework I6
  - Enter the following physician order entry dataset into Excel, and import into R.
  - Provide frequency tables, histograms, and descriptive statistics as appropriate.
  - Tabulate counts of JobCategory by Gender. Create a scatter plot of EHREntryTime vs. YearsComputerExperience. Provide boxplots of Accuracy by gender.
  - Turn in: the results of your analyses with narrative text describing the results.
  - Extra credit: turn in the R program for computing all of the above directly from the imported data frame as specified.

Homework Hint

- JobCategory
  - Attending > Resident > Intern
- One way to get R to play nice with ordinal measures is to “dummy code” them.
- You can just do this in Excel
- Extra credit: do the transformation in R