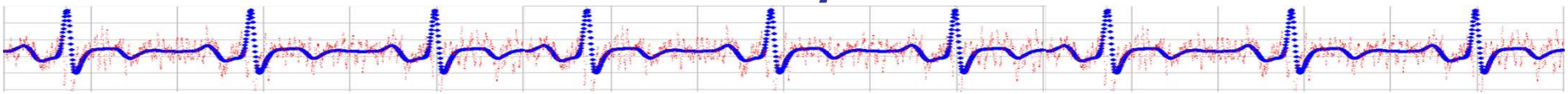
Empirical Research Methods in Information Science

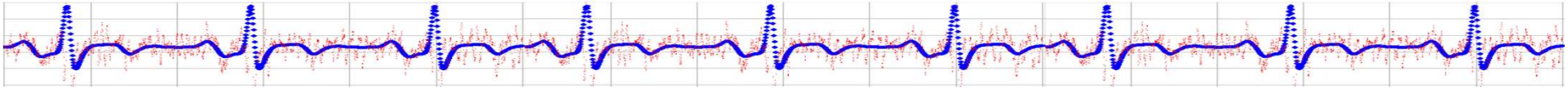
IS 4800 / CS6350



Lecture 8

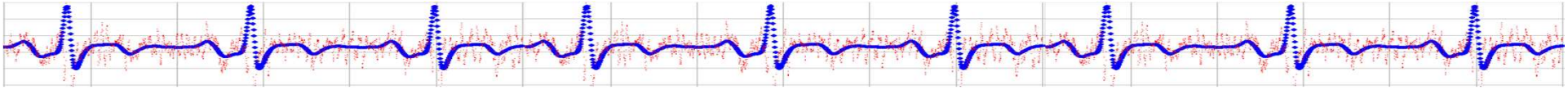
Usability and Measures

Outline



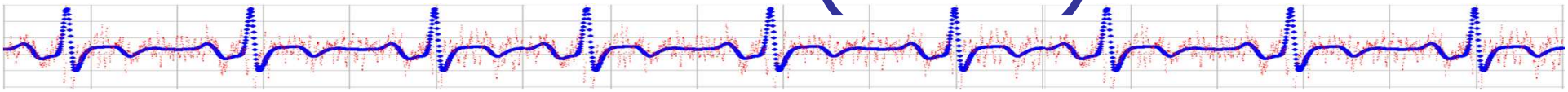
- Reading assessment
- Homework
  - Usability experiment
- Measures
  - Why worry about measures?
  - What should you consider in choosing a measure?
  - What is reliability and validity? Why care?

Homework I3



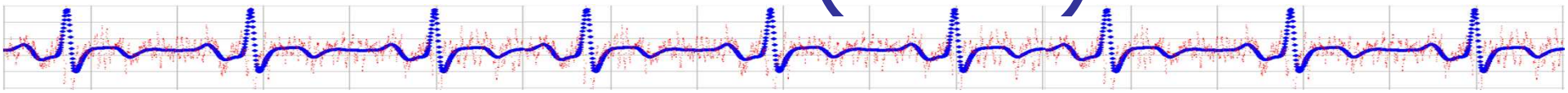
- Read example papers
- Pick an obscure piece of software with a user interface (ideally one you may have created for a class). Define two simple tasks using the software (something you can describe in 1-3 sentences and take less than three minutes to do) and write them down on two pieces of paper.

Homework I3 (cont.)



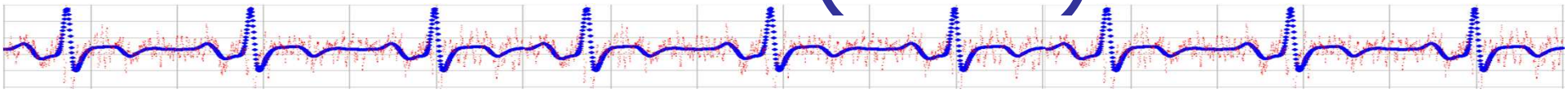
- Select two or more interval or ratio measures from pages 194-195 of the Nielsen reading that you think may be relevant to the software, in addition to at least one nominal or ordinal measure (could be sociodemographic)

Homework I3 (cont.)



- Ask three (or more) people to help you with a user study. Make sure they have not used the software before. Obtain consent using consent form. Provide a brief description of the software (but not how to use it). Then, give each participant each task and watch them attempt to complete it. **Do not provide any help**. Collect your measures.

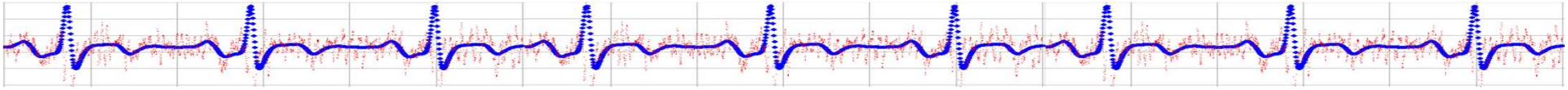
Homework I3 (cont.)



- Submit a brief writeup of your test plan, descriptive statistics of your data, and any design recommendations resulting from your tests.

- I will provide you with an example assignment, an example consent form, and the template for your consent form

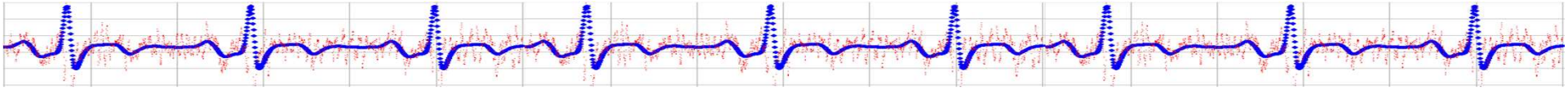
What to measure / how to measure it?



Given the choice, use a measure that…

- Is "validated"
- Has been used before in your field
- Is readily accessible or inexpensive
- That takes the least time and effort

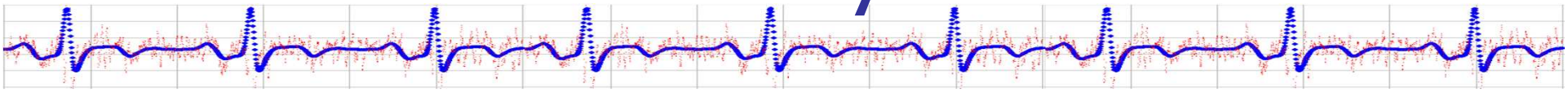
What is a validated measure?



- Has reliability
- Has validity

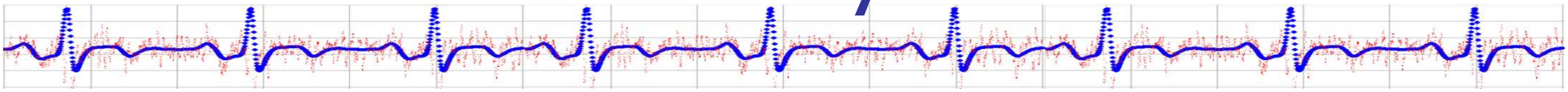
- For questionnaire measures, these are collectively referred to as a measure's "psychometrics"

Measure validity



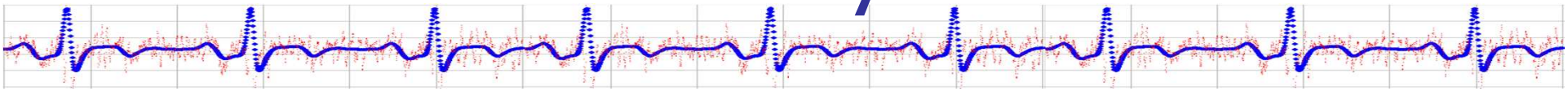
- A valid measure measures what you intend it to measure
- Carefully consider when indirectly measuring something (e.g., IQ test)
- For questionnaires, establish validity multiple ways…

Measure validity



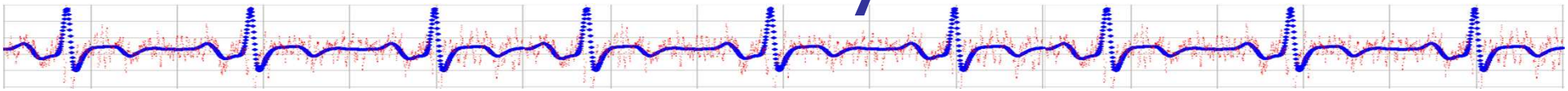
- *Face validity:* Assessment of adequacy of content; Least powerful method
- *Content validity:* How adequately does a test sample behavior get measured?

Measure validity



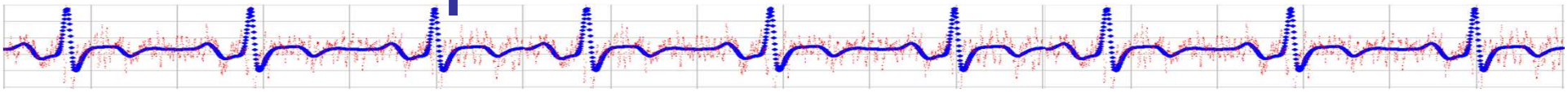
- *Criterion-related validity:* How adequately does a test score match some criterion score? Takes two forms:
  - Concurrent validity: Does test score correlate highly with score from a measure with known validity?
  - Predictive validity: Does test predict behavior known to be associated with the behavior being measured?

Measure validity

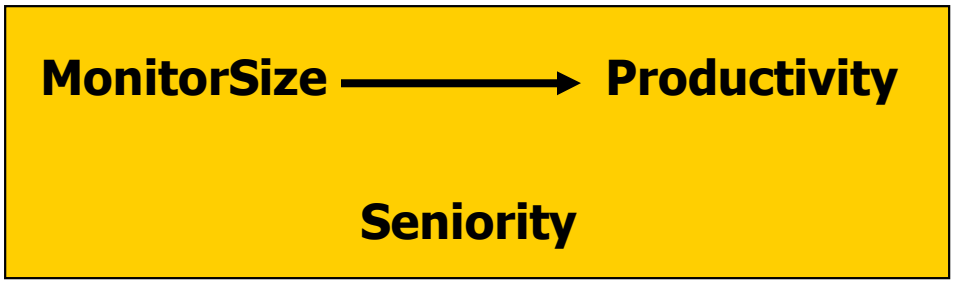


- *Construct validity:* Do the results of a test correlate with what is theoretically known about the construct being evaluated?
  - Convergent validity (subtype): measures of constructs that *should* be related to each other are
  - Discriminant validity (subtype): measures of constructs that *should not* be related are not

Example

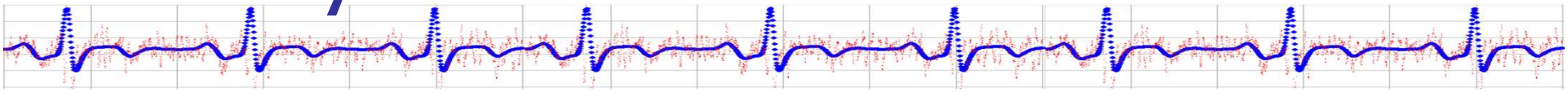


- Assume we have good evidence for this model of the world..



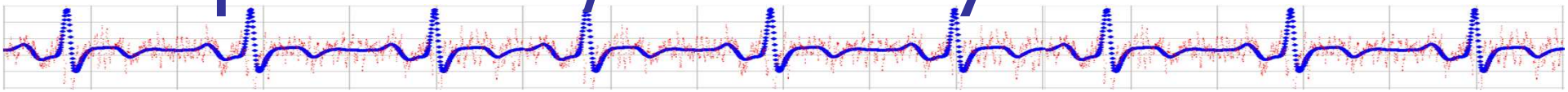
- We now propose a new measure for **Productivity**
  - What would be evidence for convergent construct validity?
  - What would be evidence for discriminant construct validity?

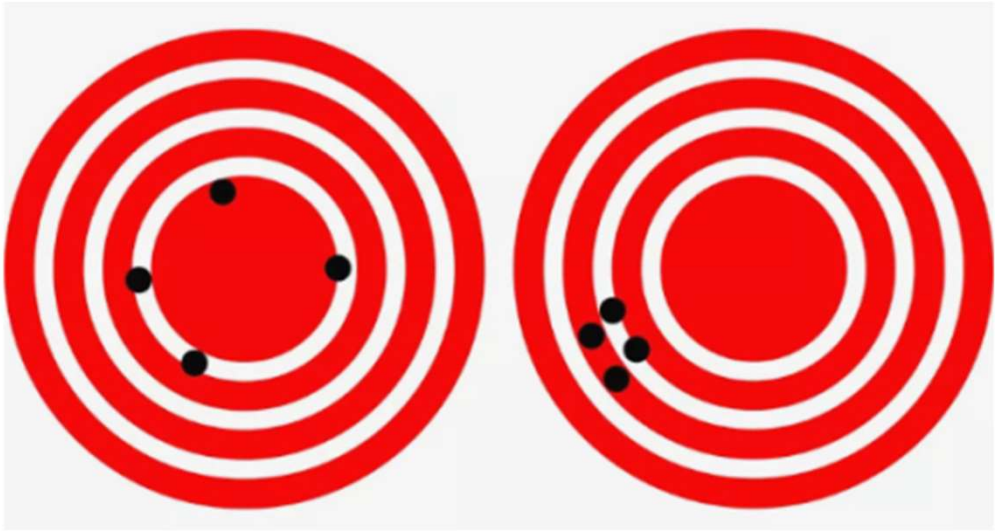
Physical measures



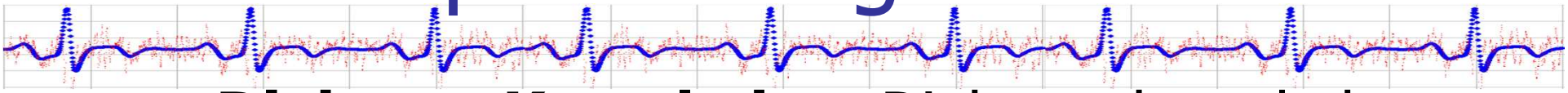
- Length, weight, time, temperature, etc.
- Validity = *ACCURACY*
  - An accurate measure produces results that agree with a known standard (i.e., is "correct")

Accuracy vs. precision/reliability



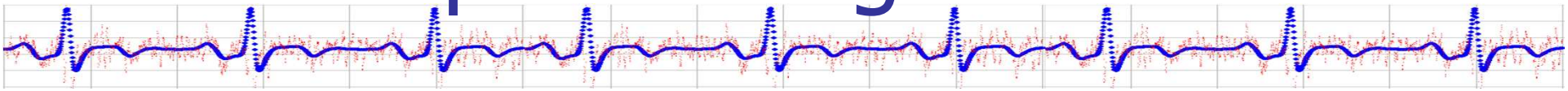


Example: How good is it?



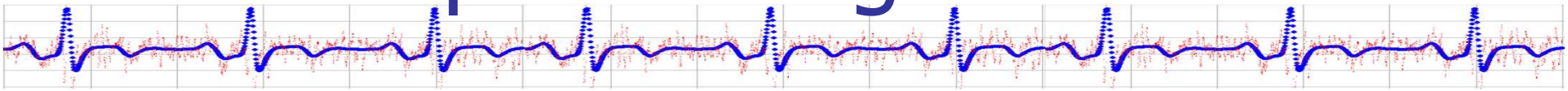
Diabetes Knowledge. Diabetes knowledge will be assessed using the Diabetes Knowledge (DKN) Scales, three separate 15-item multiple choice questionnaires that measure general diabetes knowledge. Reliability for the items in the scales (Cronbach's alpha) was 0.92, indicating high internal consistency. 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?



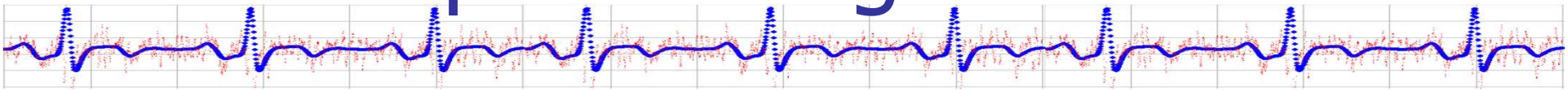
Fitness & Mobility will be assessed using timed maximal walking velocity. This measure, already assessed routinely for all patients, involves having subjects walk along an 11-meter, straight, flat walkway as fast as possible. Each subject will have three trials, with 30-second intervening rest periods. 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?



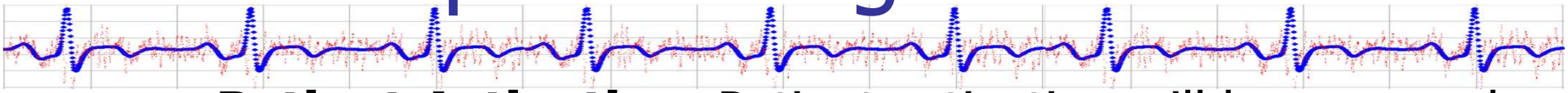
Loneliness will be assessed using the UCLA Loneliness Scale. This measure is highly reliable, both in terms of internal consistency (alpha ranging from .89 to .94) and test-retest reliability over a 1-year period (r = .73). Convergent construct validity for the scale was indicated by significant correlations with other measures including 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?



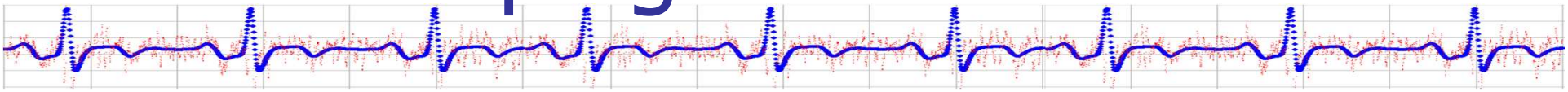
Exercise Self Efficacy. The five-item Self Efficacy Scale for exercise assesses perceived confidence to perform exercise across a wide variety of challenging situations. Recently, a new measure was developed addressing the multidimensionality of the self-efficacy construct. The short form (α = .82) of 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 ($r=-.07$, $p<.01$).

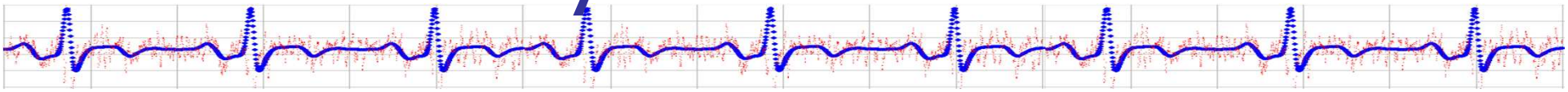
Developing a new measure



- 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*

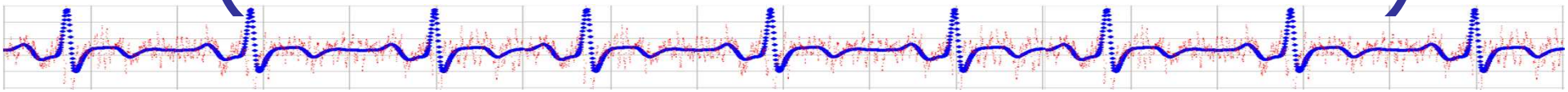
- How would you validate this measure?

Questionnaire validation: Summary

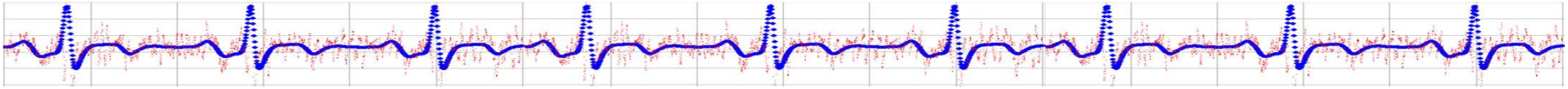


- Reliability
  - Test-retest
  - Internal consistency
- Validity
  - Face
  - Content
  - Criterion-related
    - Concurrent
    - Predictive
  - Construct
    - Convergent
    - Discriminant

Scales of measurement (aka levels of measurement)

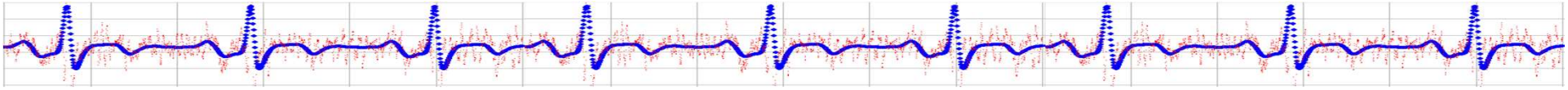


Scales of measurement



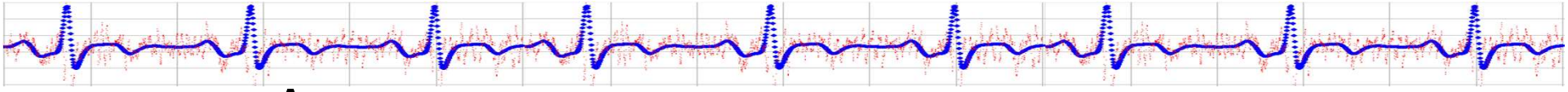
- *Nominal Scale*
  - Lowest scale of measurement involving variables whose values differ by category (e.g., male/female)
  - Values of variables have different names, but no ordering of values is implied
- *Ordinal Scale*
  - Higher scale of measurement than nominal scale
  - Different values of a variable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

Scales of Measurement



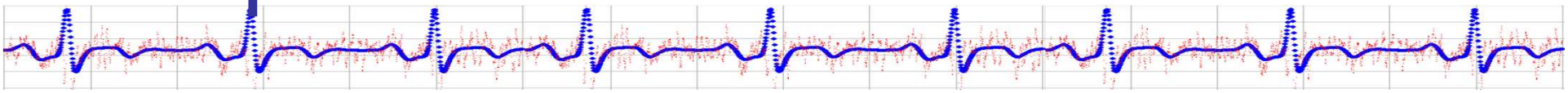
- *Interval Scale*
  - Scale of measurement on which the spacing between values is known (e.g., IQ)
  - No meaningful zero point
- *Ratio Scale*
  - Similar to interval scale, but with a true zero point (e.g., number of lever presses)

What kind is it?



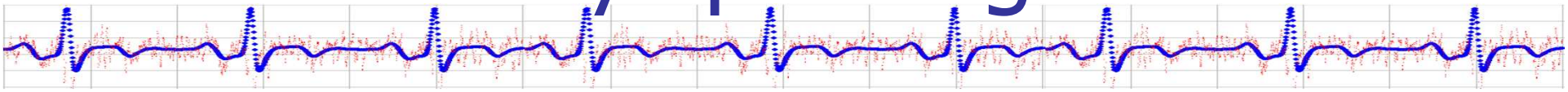
- Age
- Gender
- Job Category (Engineer, Manager…)
- Weight
- School Year (Freshman…)
- Temperature (Celsius)
- Olympic medal (Gold, Silver, Bronze)
- Monitor Size
- Weather (Rain, Snow, …)
- Salary
- Productivity (wpd)
- Owns Pet (or not)

A final word on scale item questionnaires



- Treat a single item as ordinal
- Treat a composite questionnaire (with at least six items) as interval
- Will discuss rationale later…

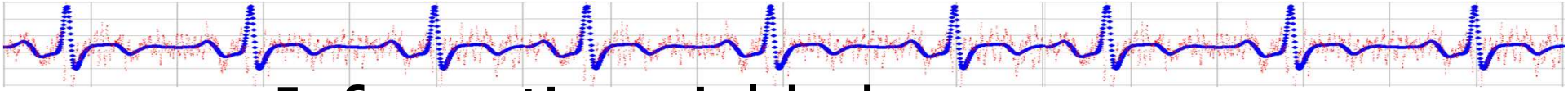
Practically speaking



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

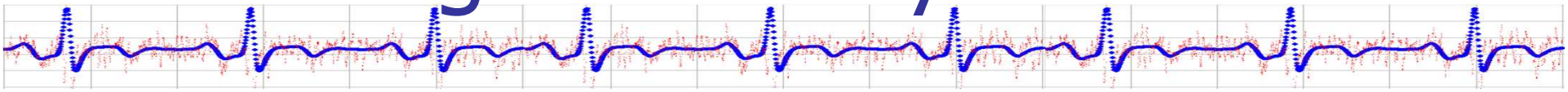
  And
  - Histogram (later)

Factors affecting your choice of a scale of measurement



- Information yielded
  - Nominal scale: least information
  - Ordinal scale: adds more information
  - Interval and ratio scales: most information
- Statistical tests available
  - Nominal and ordinal data (nonparametric) tests less powerful than interval and ratio data (parametric) tests
  - Use the scale that allows most powerful statistical test

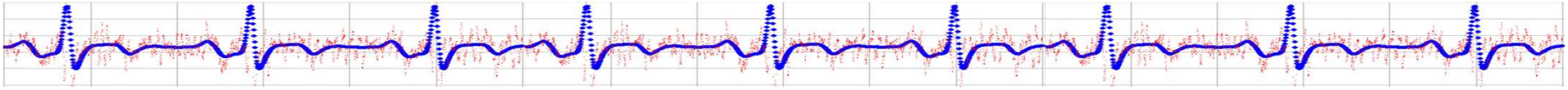
Ecological validity



- The degree to which a measure corresponds to what happens in the real world.
- Example:

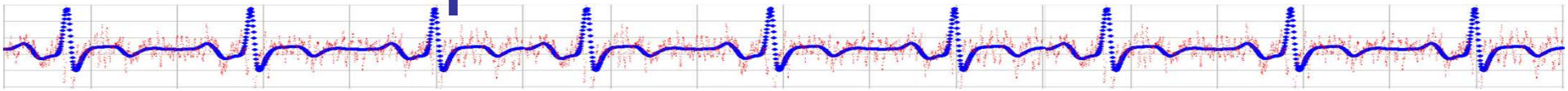
  Assessing productivity/day in the lab vs. assessing productivity/day in the office

Concerns with measures



- Sensitivity
  - Is a dependent measure sensitive enough to detect the change you are interested in?
  - An insensitive measure will not detect subtle behaviors
- Range Effects
  - *Ceiling effect:* When a dependent measure has an upper limit
  - *Floor effect:* When a dependent measure has a lower limit

Example

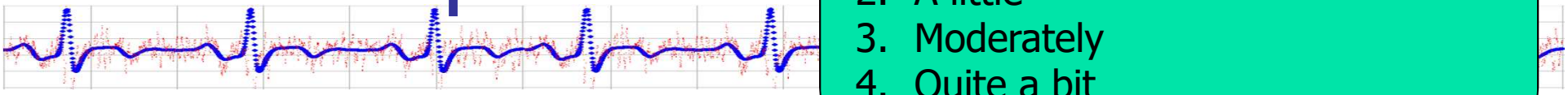


- You want to assess the effect of TV viewing on whether people are happy or not (yes/no)
- You run an experiment in which participants are randomized to watch either 2 hrs or 0 hrs of TV per day for a week, then answer your question.

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

- What's going on?

Example



At the time of the beep, I felt _____ interested.

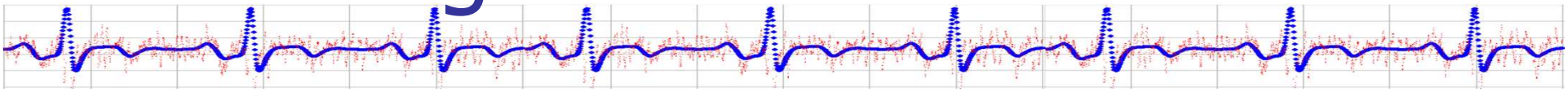
1. Very slightly or not at all
2. A little
3. Moderately
4. Quite a bit
5. Extremely

- You want to assess the e
  positive affect, measured on a 1-5 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	5.0
2	No TV	4.7
3	TV	4.9
4	No TV	5.0

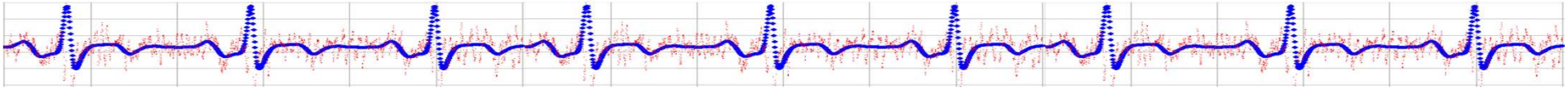
- What's going on?

Ceiling and floor effects



- Affect data in two ways
  - Limiting values of your highest or lowest data point
  - Variability of scores within affected treatments is reduced
    - May cause misleading results from statistical analysis of data

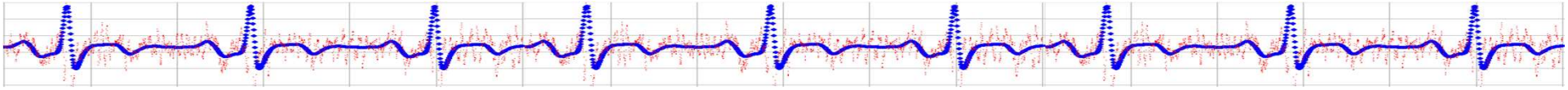
Some special types of measures



Behavioral Measure

- Record actual behavior of subjects
- 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 in a future class

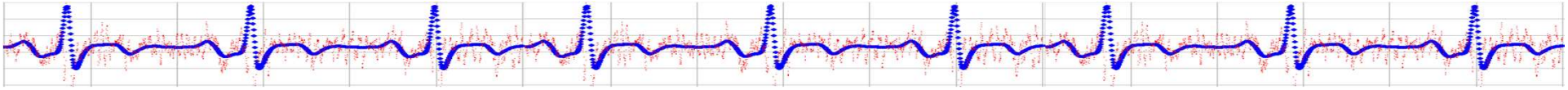
Some special types of measures



Physiological measures (sensors?)

- 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 a subject's body (e.g., arousal)
- Must infer psychological states

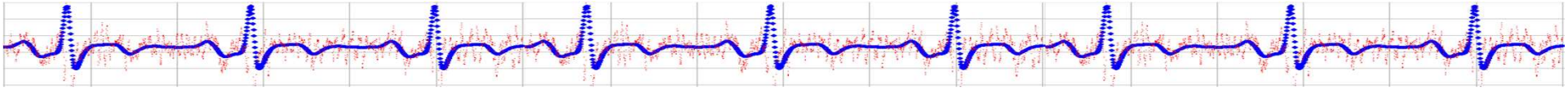
Some special types of measures



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 a person's attractiveness 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

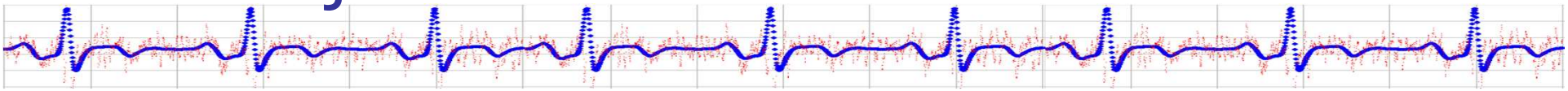
Some special types of measures



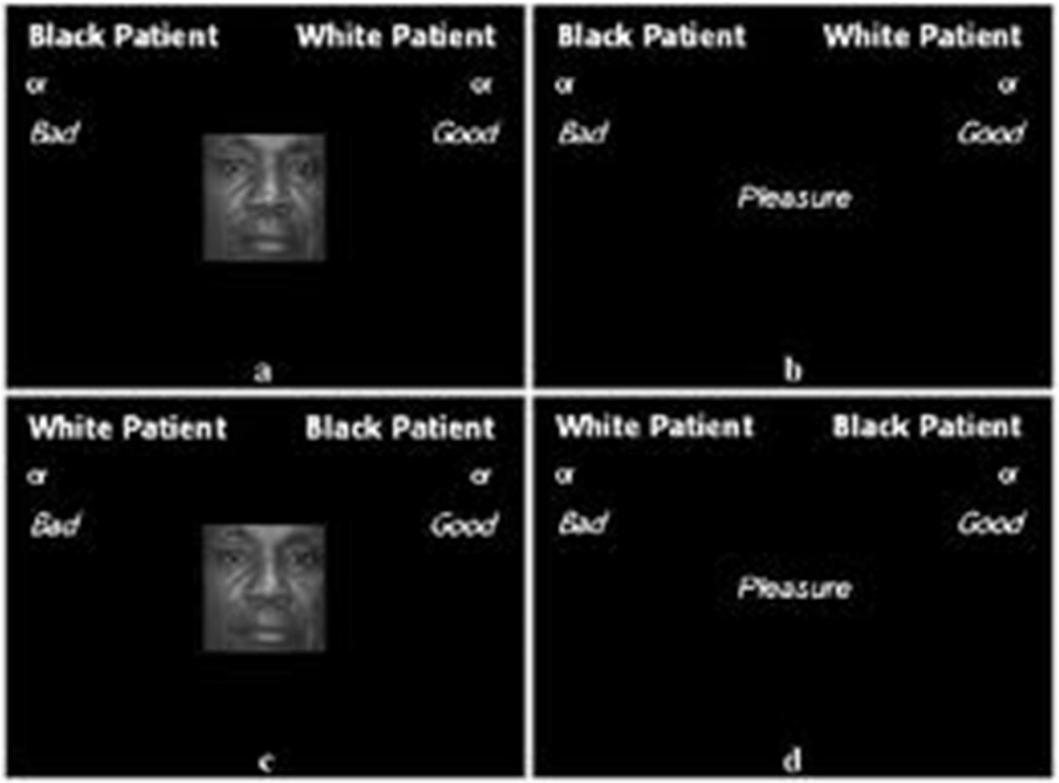
- Physical measures (sensors?)
  - Temperature
  - Pressure
  - ?
- System measures ("sensors"?)
  - Profiling (%use, %CPU, etc.)
  - Runtime (clock or CPU)
  - Mean time between failures (MTBF)
  - ?

Implicit measures

Subject is unconscious of measurement

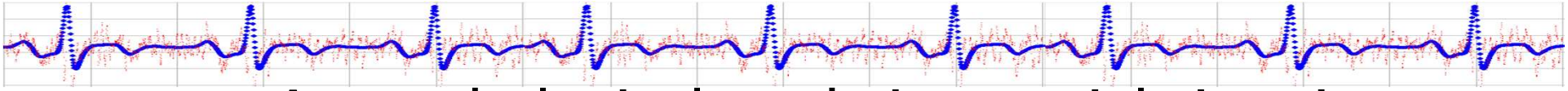


- Uses rapid, unconscious categorization task to tease out biases
- Assume quicker reaction times are associated with stronger concept associations



https://implicit.harvard.edu/implicit/takeatest.html

Reactivity in psychological research

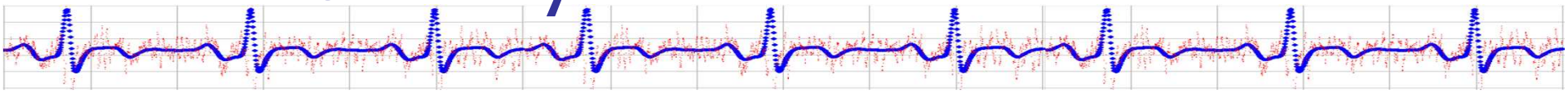


- A psychological study is a social situation
- A participant's social history can affect how he or she responds to a study
- You should not assume that your participant is a passive recipient of the parameters of your study
- Simply observing someone changes his or her behavior

In the robotic vacuum study from last time, how might reactivity manifest?

What could you do about it?

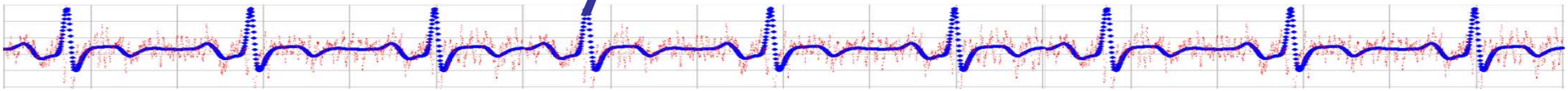
Reactivity



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

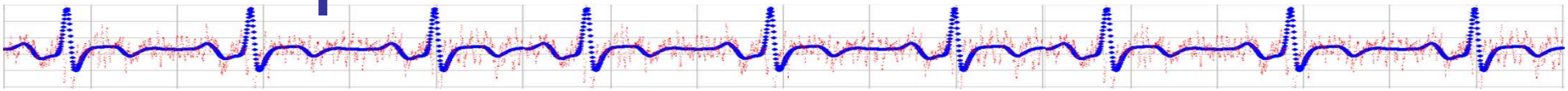
Reactivity



Role attitude cues (attitude adopted by a participant) can affect outcome of a study

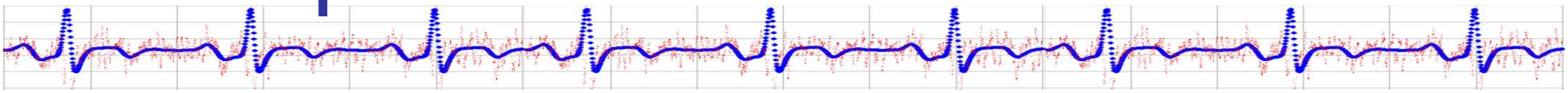
- Cooperative attitude: Participant wants to help researcher
- Defensive or apprehensive attitude: Participant is suspicious of experimenter and situation
- Negative attitude: Participant motivated to ruin a study

Experimenter effects



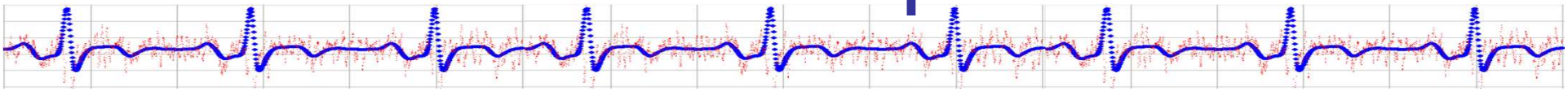
- An experimenter can unintentionally affect how a participant behaves in a study
- Experimenter bias occurs when the experimenter's behavior influences a participant's behavior
  - Two sources of experimenter bias
    - *Expectancy effects:* An experimenter expects certain types of behavior from participants, 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 effects



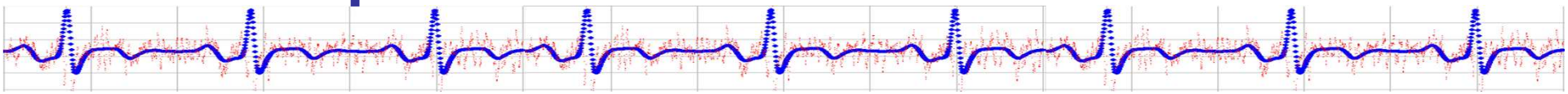
- Experimenter bias affects internal and external validity
- Steps to reduce:
  - *Blind technique*: the experimenter or subject does not know the condition to which a participant has been assigned
  - *Double-blind technique:* neither the experimenter nor participant knows the participant's assignment condition
    - Automate the experiment

Additional concepts in Ch 5



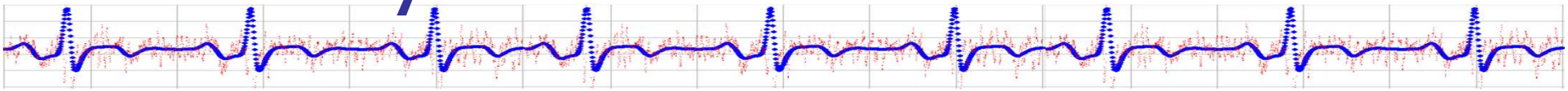
- Pilot Study
  - E.g. for Sample Research Plan
- Manipulation Check
  - Test if IV had intended effects on your participants (e.g., could they tell the difference; did they perceive experiment as you thought?)
  - E.g., Mouse frustration study (Klein)

Chapter 13



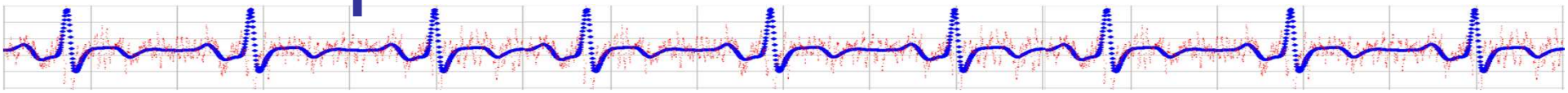
Describing Data

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…

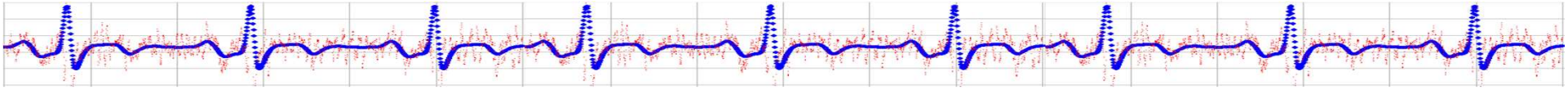
Steps involved in EDA



1. Enter your data
2. Display frequency distributions on a histogram for each measure

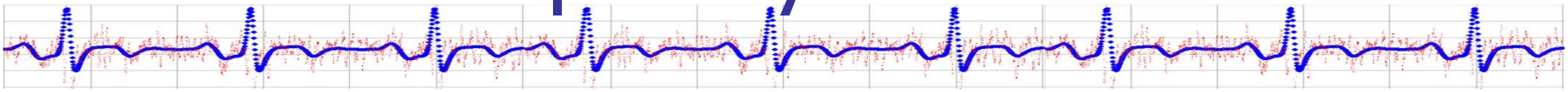
   For non-nominal measures: examine graphs for normality and skewness
3. Graph data (bar graph, line graph, or scatterplot) so that you can visually inspect relationships

Stacked vs. unstacked data table format?



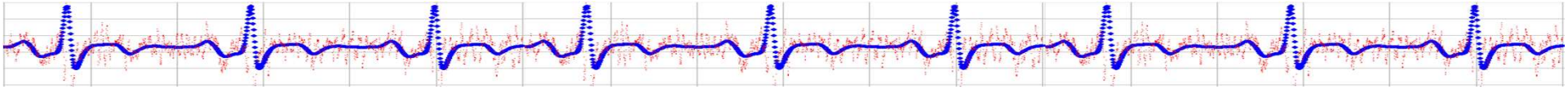
- Unstacked = 1 row per subject
- Stacked = 1 row per observation

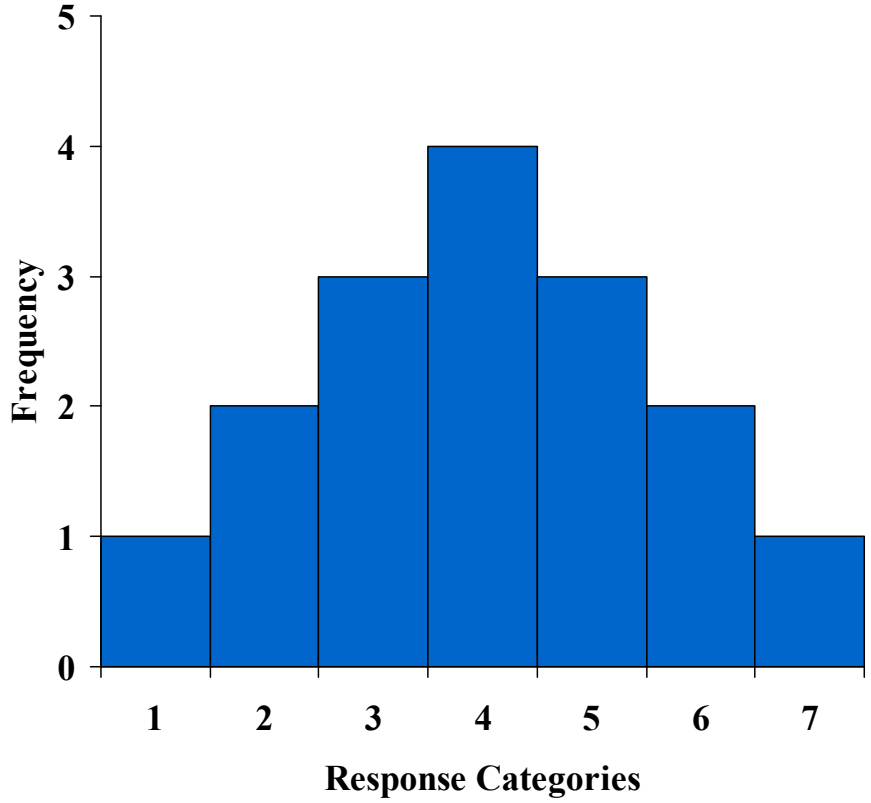
The frequency distribution



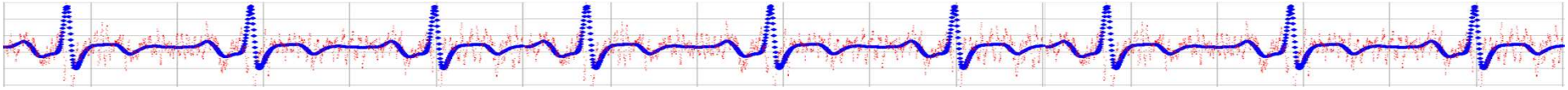
- Represents a set of mutually exclusive categories into which actual values are classified
- Can take the form of a table or a graph
- Graphically, a frequency distribution is shown on a *histogram*
  - A bar graph in which the bars touch
  - The y-axis represents a frequency count of the number of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

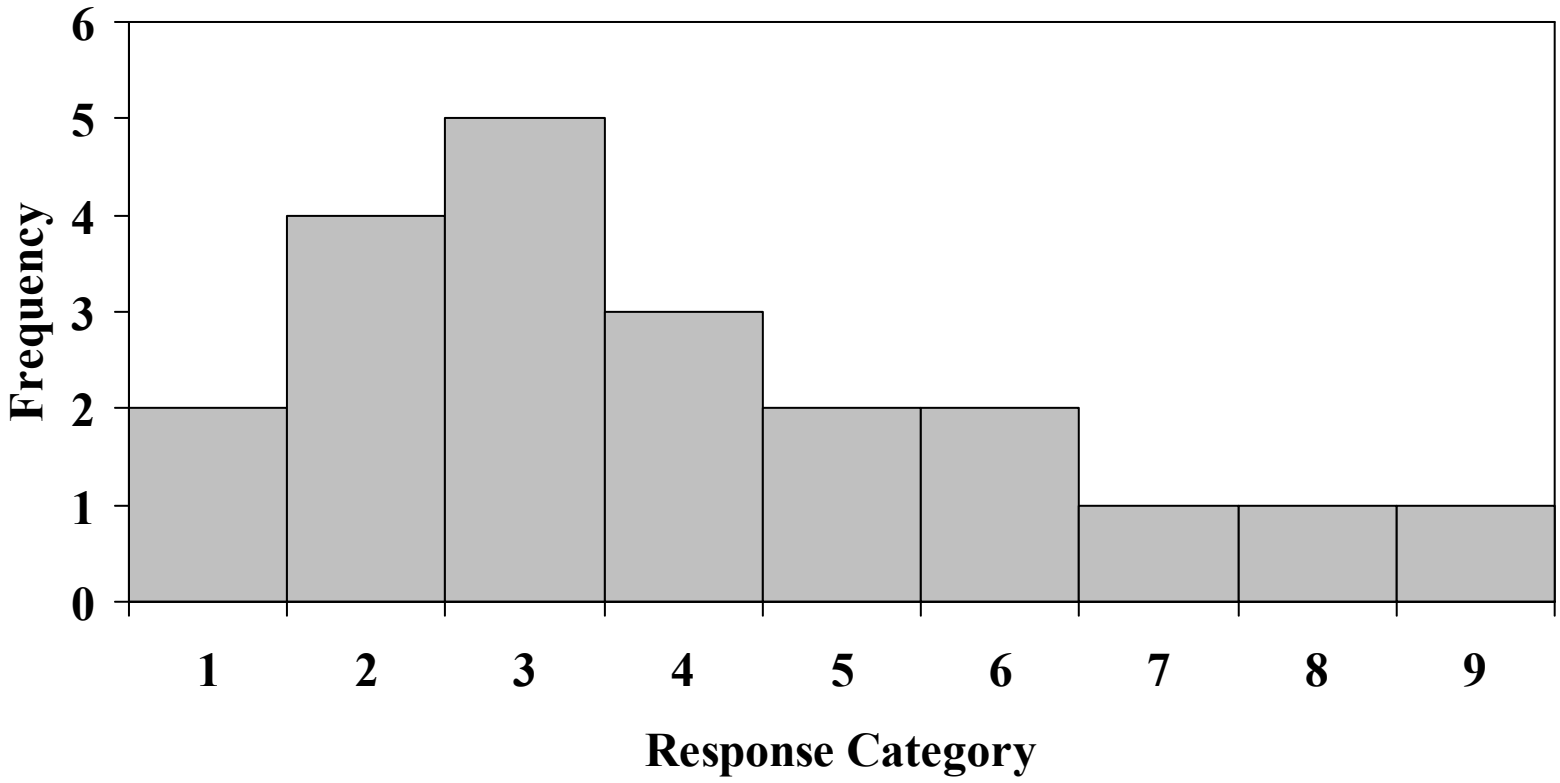
Histogram showing a normal distribution



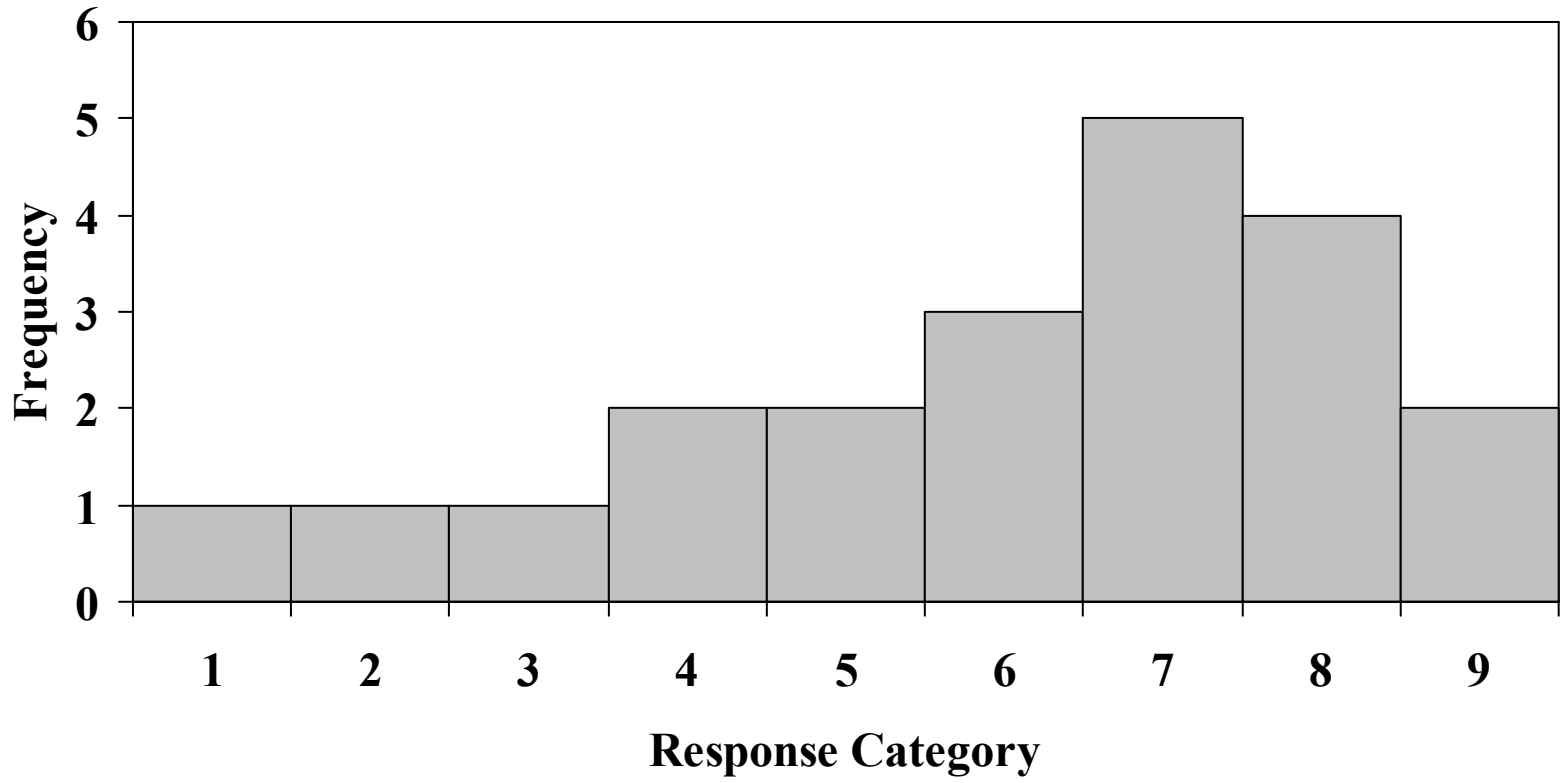


Histogram showing a positive skew

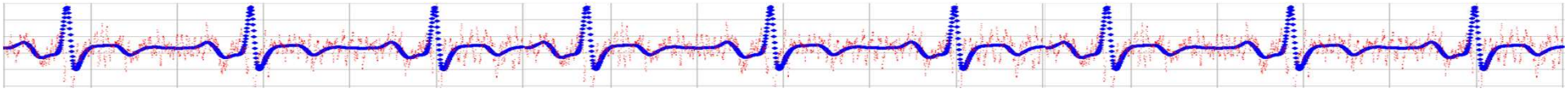


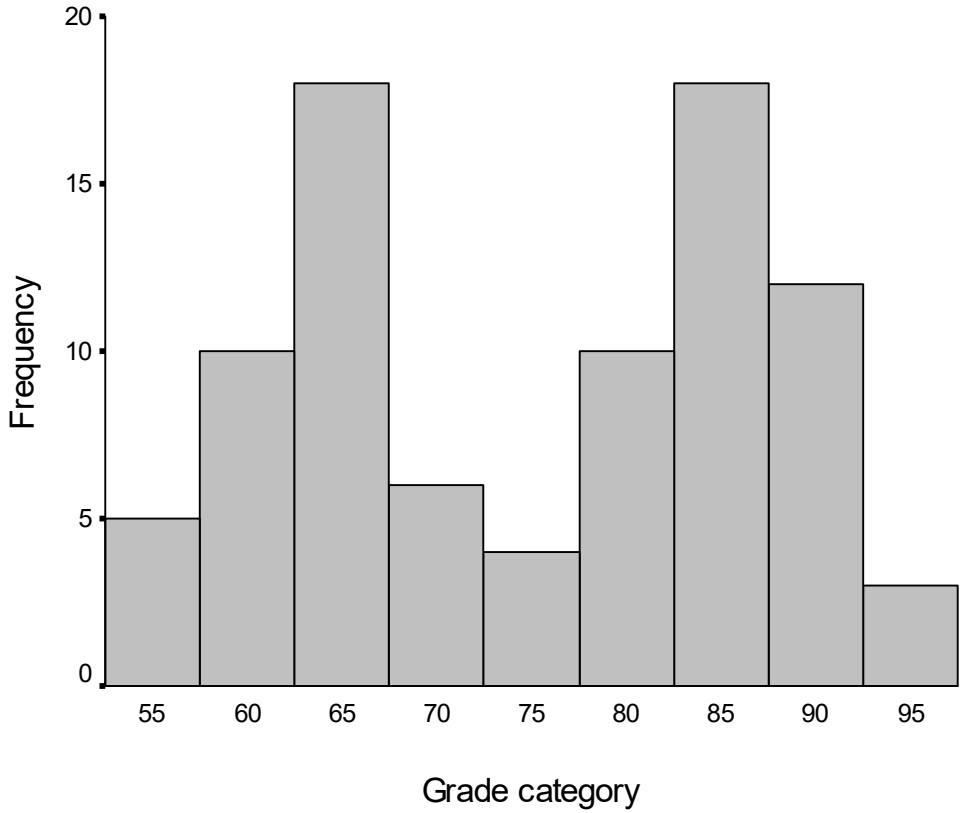


Histogram Showing a Negative Skew

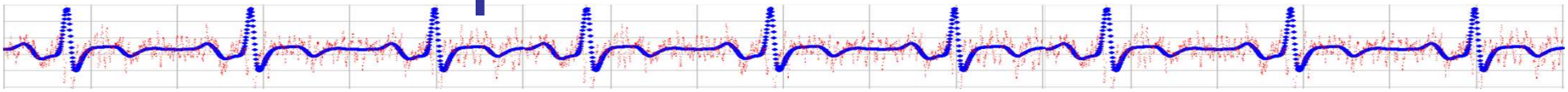


A bimodal distribution



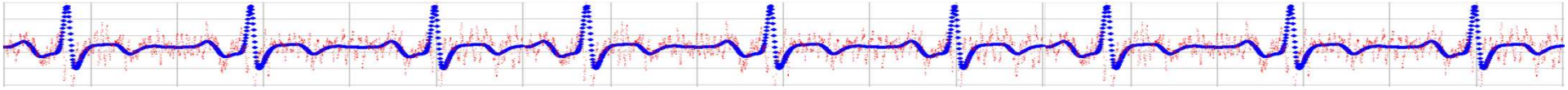


Descriptive statistics



- Statistic = a number used to describe some feature of a group of measurements
- Class rule: For every measure you must have
  - Exactly one statistic describing a measure of center
  - Zero or one statistic describing a measure of spread

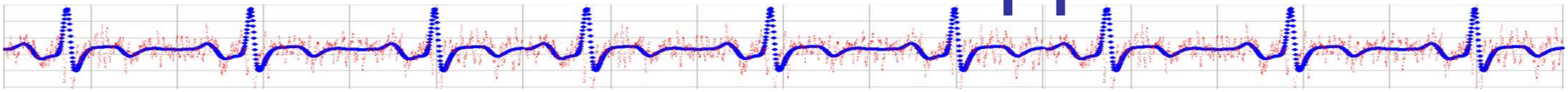
Measures of center



- Mean
- Median
- Mode

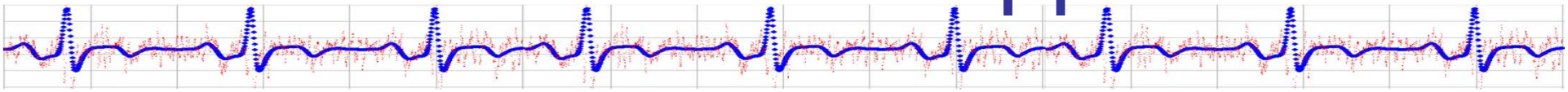
- Whazzit?
- When to use?

Measures of center: Characteristics & applications



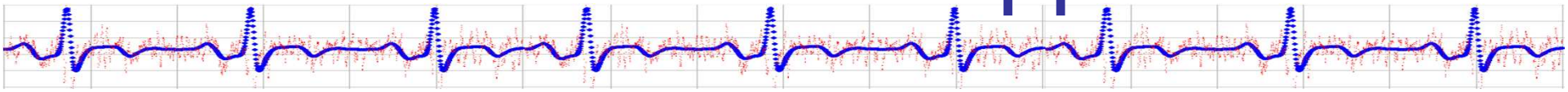
- *Mode*
  - Most frequent score in a distribution
  - Simplest measure of center
  - Scores other than the most frequent not considered
  - Limited application and value

Measures of center: Characteristics & applications



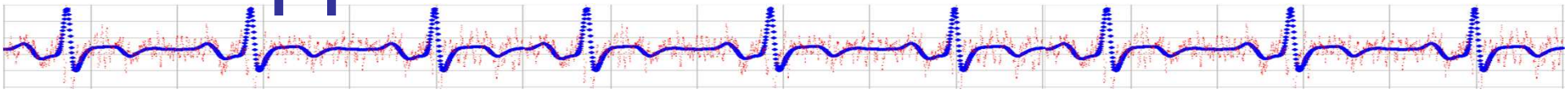
- *Median*
  - Central score in an ordered distribution
  - More information taken into account than with the mode
  - Relatively insensitive to outliers
  - Prefer when data is skewed
  - Used primarily when the mean cannot be used

Measures of center: Characteristics & applications



- *Mean*
  - Average of all scores in a distribution
  - Value dependent on each score in a distribution
  - Most widely used and informative measure of center

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 or ratio data do not meet requirements for using the mean (skewed but unimodal), or if significant outliers