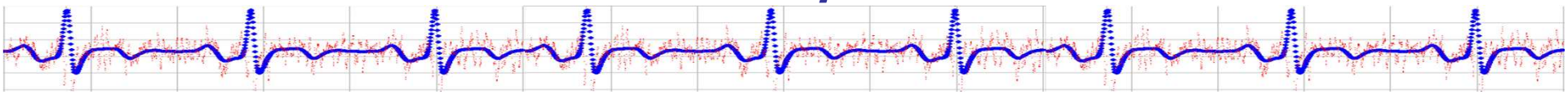
Empirical Research Methods in Information Science

IS 4800 / CS6350

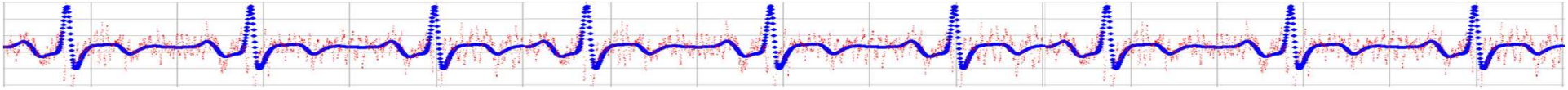


Lecture 10

Survey design

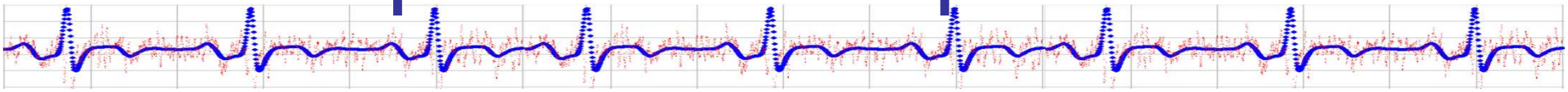
Maybe start hypothesis testing

Outline



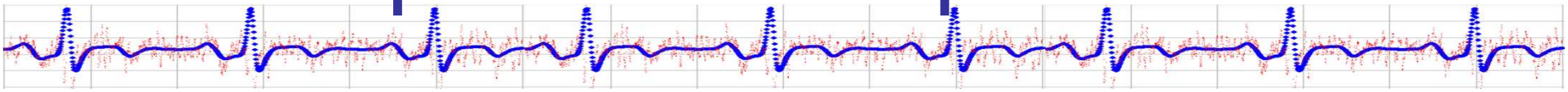
- Reading assessment
- Homework I4 – plan
- Survey/instrument design
- Maybe start hypothesis testing

Homework I4: Design a composite self-report measure



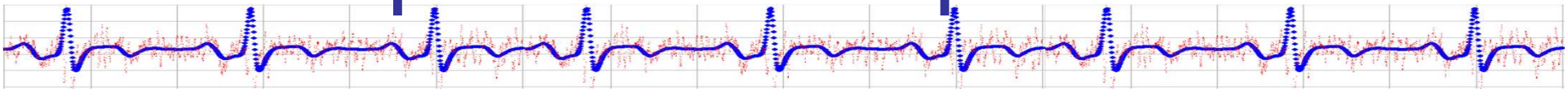
- Design a new composite self-report measure (e.g. "homework procrastination") … but your own idea
- Assume it only has one factor, but use at least five scale items
- Incorporate information from at least one literature reference

Homework I4: Design a composite self-report measure



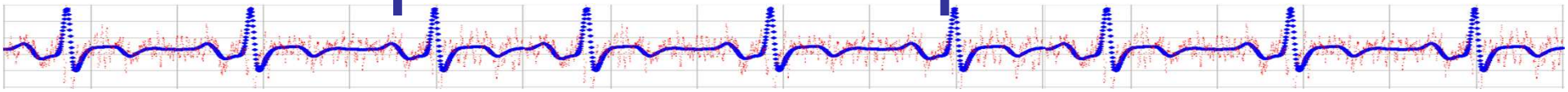
- Assess the face and content validity of your measure and work through a bivariate analysis of your items
- Implement questionnaire on surveymonkey.com or Google forms

Homework I4: Design a composite self-report measure



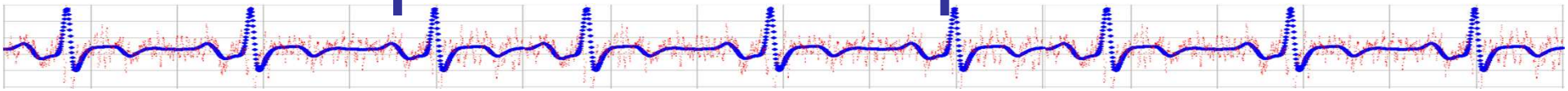
- Decide on one method for assessing validity (besides face & content) for your measure that you can also assess in a self-report questionnaire. This should be an additional question (or an additional previously validated composite measure) on your survey and should provide a numeric measure

Homework I4: Design a composite self-report measure



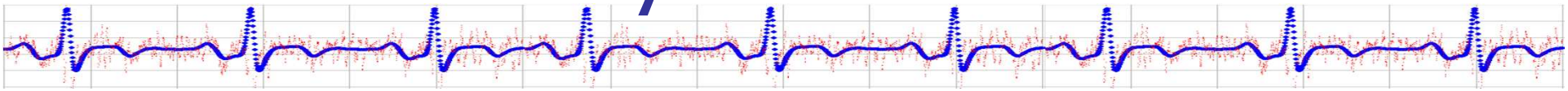
- Post your questionnaire on Piazza
- You are obligated to reply to any questionnaires posted within 48 h!
- Compute the reliability (internal consistency) of your measure using Python
- Compute descriptive statistics for your measure and any other items you may have included on the questionnaire

Homework I4: Design a composite self-report measure



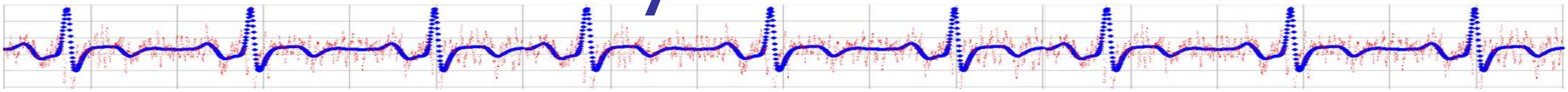
- Assess the validity of your measure (you can do this qualitatively, e.g., using scatterplots)
- Document and submit all of the above
- You may work individually or in teams of two
- Due 2/20

Meta-analyses



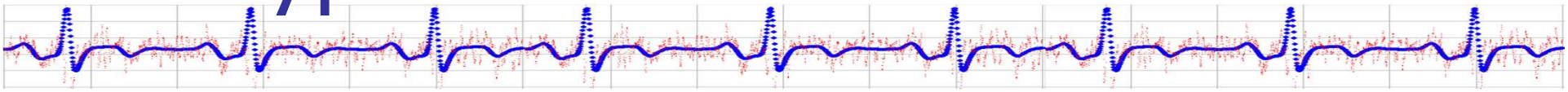
- Compare/integrate "all" studies that have investigated a given phenomena
  - E.g., use of a particular medication for a particular disease
- Common in the literature (esp. medical)
- Very methodical
  - Search for articles
  - Eligibility criteria
  - Statistical analyses

Meta-Analysis



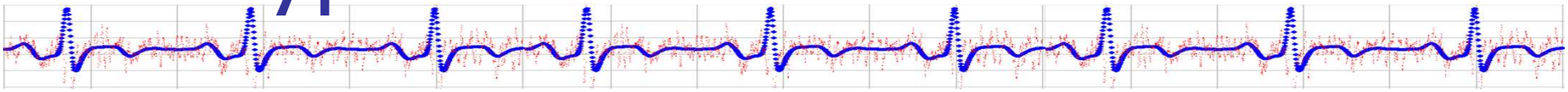
- New terms(?)
  - Level of Significance
  - Effect Size
  - Type I & II errors

Type I error



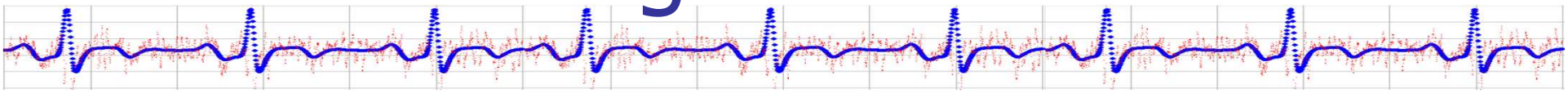
- Rejection of a true null hypothesis (also known as a "false positive" finding)
- Often represented by the Greek letter alpha (α)

Type II error



- Failure to reject a false null hypothesis (also known as a "false negative" finding)
- Often represented by the Greek letter beta **(**β **)**

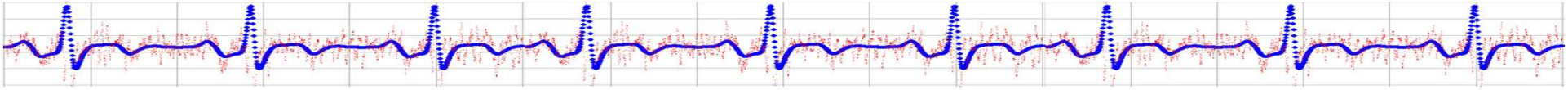
Level of significance



Probability of rejecting a null hypothesis by the test when it is really true, which is denoted as α. That is, P(Type I error) = α.

The level of significance 0.05 is related to the 95% confidence level

Power



- Probability that a test will reject the null hypothesis when it is, in fact, false.
- 1 – β (type II error rate)
- High power is desirable. Like β, power can be difficult to estimate accurately, but increasing the sample size always increases power

Type I and Type II errors

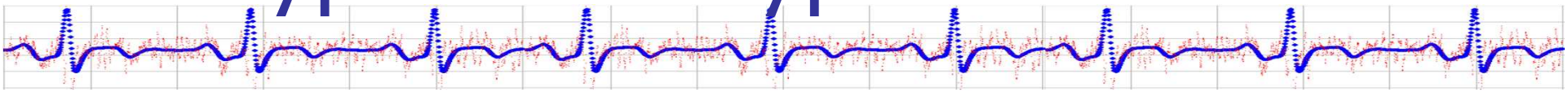
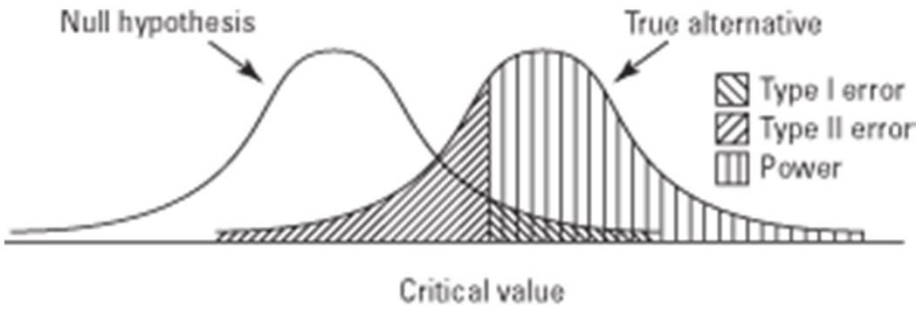
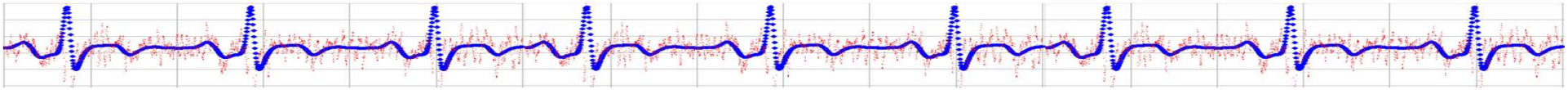


Table 1. Types of Statistical Errors

	H_0 is actually:	
	True	**False**
Reject H_0	Type I error	Correct
Accept H_0	Correct	Type II error

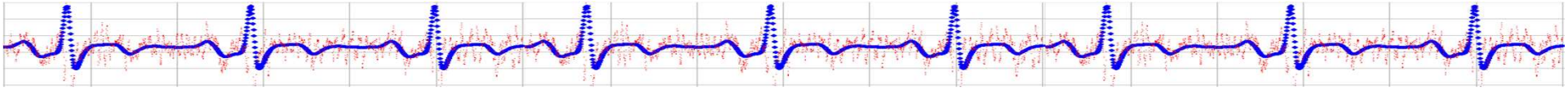


Effect size



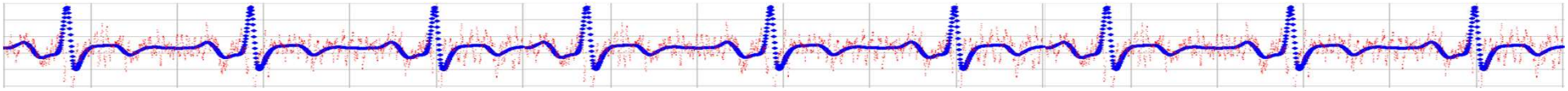
- Measures strength of the relationship between two variables on a numeric scale
- E.g., Data on height (men/women); notice men usually taller; the difference between the heights is effect size (greater effect size -> greater difference)

Effect size

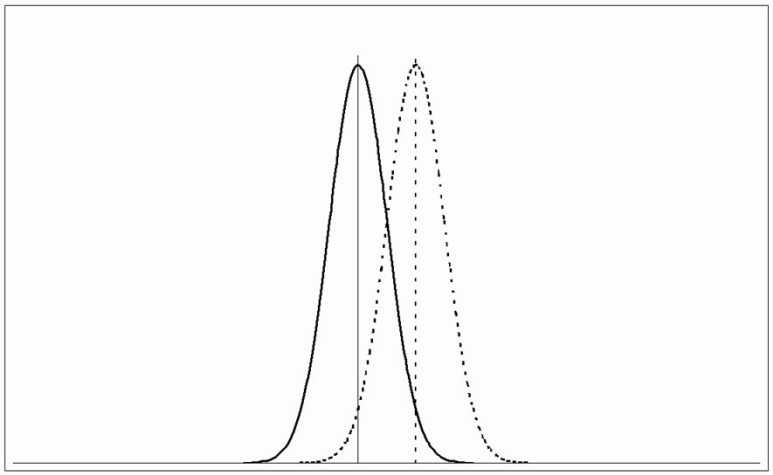


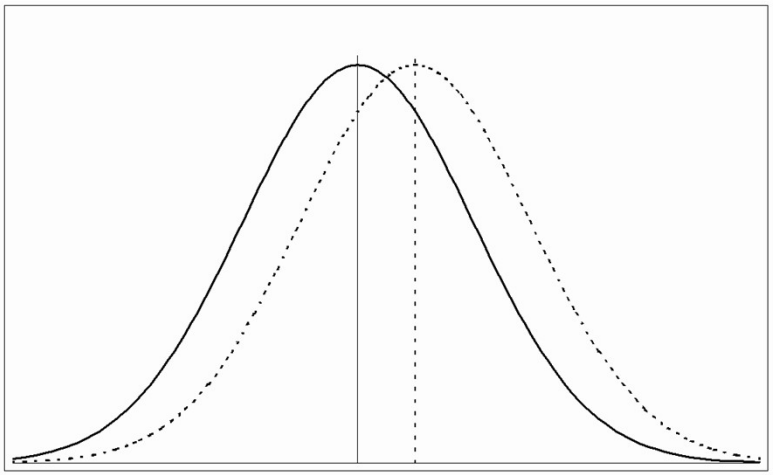
Effect size is usually measured in one of three ways: (1) standardized mean difference, (2) odd ratio, (3) correlation coefficient

Effect size mean

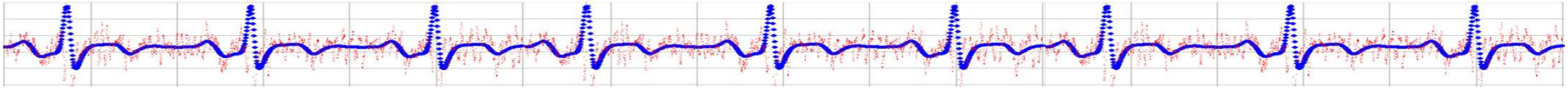


$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$





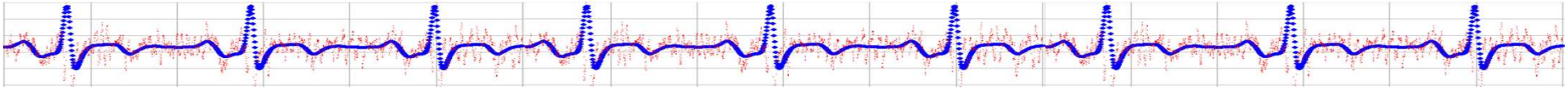
Effect size: correlation



Estimate the amount of the variance within an experiment that is "explained" or "accounted for" by the experiment's model

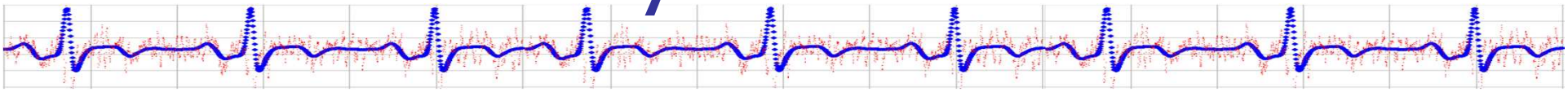
Can use Pearson’s correlation

Pearson's correlation



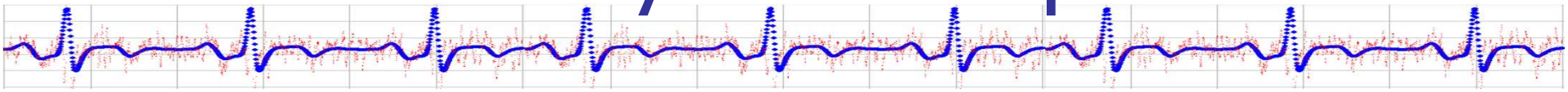
- r for effect size
- -1 to 1
- Guidelines:
  - Small 0.1
  - Medium 0.3
  - Large: 0.5

Meta-Analyses



- Effect Size
  - Measure of how much difference exists between treatment groups in an experiment
  - How to assess as common metric?
    - E.g., compare effect of large monitors on productivity
      - Study 1 measures widgets per day
      - Study 2 measures subjective assessment of managers
  - How to integrate across studies?

Meta-analysis example



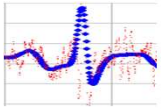
CHI 2007 Proceedings • Faces & Bodies in Interaction April 28-May 3, 2007 • San Jose, CA, USA

A Meta-Analysis of the Impact of the Inclusion and Realism of Human-Like Faces on User Experiences in Interfaces

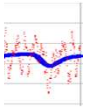
Nick Yee, Jeremy N. Bailenson, Kathryn Rickertsen
Department of Communication
Stanford University, Stanford, CA
{nyee, bailenson, kathrynr}@stanford.edu

METHOD

Selection of Studies



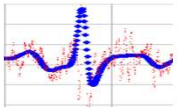
The studies considered for inclusion in this analysis were culled from bibliographic indexes related to the fields of psychology, computer-mediated communication (CMC), and virtual reality. These included Expanded Academic ASAP, Google Scholar, Google keyword, PsycInfo, PsycArticles Fulltext Search, InterDok, ProQuest, and SearchPlus. In this initial pass, articles that appeared to report an experimental study of anthropomorphism, embodied agents, or agent realism were collected and reviewed. Sources were only considered if they were published in a peer-reviewed journal or in published conference proceedings. This ensured a basic level of



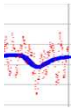
The literature review yielded 106 studies. Several selection criteria were then applied. First, an article was included only if it was an experimental study that manipulated the variables of interest and contained clear reports of quantitative data relating to the outcome of different conditions. Thus, purely qualitative studies involving open-ended self-reports or observational user studies without quantitative coding schemes or dependent variables were removed.

Of these 25 studies, the average year of publication was 2001.96 (SD = 2.29) with a median of 2002. The average sample size within each study was 45.40 (SD = 35.55). With regard to study location, 13 were conducted in the US or Canada, 9 were performed in Europe, and the remaining 3 were conducted in Asia. And finally, with regard to equipment used, 17 were conducted on desktop equipment, 6 were conducted using immersive virtual reality, and the remaining 2 were conducted on a large projected screen.

Effect Size Calculations

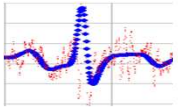


To generate the necessary effect size tabulations in order to test our hypotheses, we tabulated several possible effect sizes for each paper depending on the available conditions. First, we tabulated the results of performance data separately from the results of subjective data. Performance data might include time to task completion, accuracy measures, or similar behavioral measures. Subjective data, on the other hand, was any measure that was based on self-report or survey data. Second, we tabulated effect sizes based on two kinds of comparisons between conditions. We

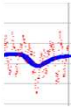


RESULTS

Formal Meta-Analyses



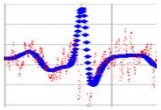
The results of the effect size and significance value aggregation are listed in Appendix A for each individual study and the overall values. The overall effect sizes of the four comparison conditions ranged from -.04 to .14. While three of the four comparison conditions were highly significant at p levels of less than .05, the comparison of high-low realism using performance measures was not significant, with $p = .14$.



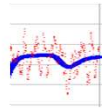
APPENDIX A – EFFECT SIZES AND SIGNIFICANCE VALUES OF STUDIES INCLUDED

	Performance		Subjective		
	Face vs. No Face	High vs. Low Realism	Face vs. No Face	High vs. Low Realism	N
Okonkwo & Vassileva, 2001 [41]		r = 0, z = 0.24		r = 0.03, z = 0.84	12
Moundridou, Virvou 2002 [37]	r = 0.1, z = 0.39		r = 0.48, z = 4		48
Hongpaisanwiwat & Lewis, 2003 [23]	r = 0, z = -0.02	r = 0.07, z = 0.45			50
Burgoon, Bengtsson, Bonito, Ramirez, & Dunbar, 1999 [11]	r = 0.03, z = 0.2	r = -0.03, z = -0.17	r = 0, z = -0.04	r = 0.12, z = 0.8	50
Bailenson, Beall, & Blasovich, 2002 [2]			r = 0.51, z = 1.92	r = 0.16, z = 0.46	30
Burgoon, Bonito, Bengtsson, Cederberg, Lundeberg,					

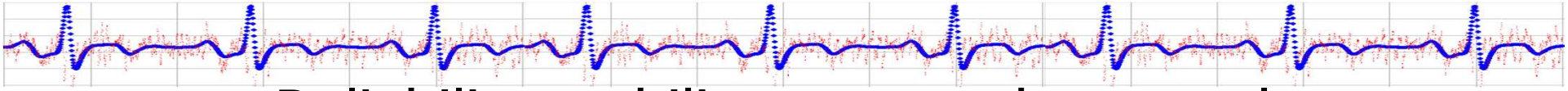
- Notes:
  - r is a measure of effect size; r^2 is the amount of variance in the DV accounted for by the IV.



In our meta-analysis, we had also separated: 1) studies that compared interacting with an agent that had no facial representation versus an agent that had a facial representation (i.e., the yes-no comparisons), and 2) studies that compared interacting with faces of low realism versus faces of high realism (i.e., the high-low comparison). A comparison of these two groups of effect sizes revealed that the effect sizes from yes-no comparisons (n = 25, r = .16) were significantly larger than those from the high-low comparison (n = 18, r = .07), z = 2.43, p = .02.

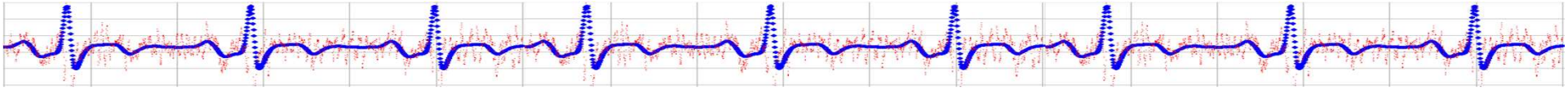


Notes



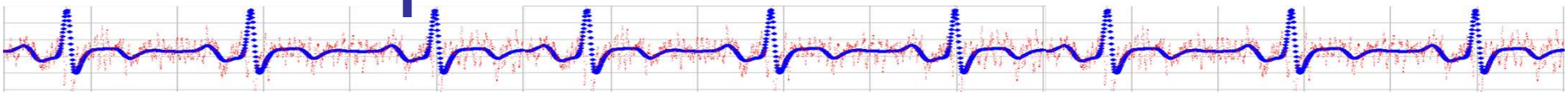
- Reliability – ability to reproduce results
- Accuracy – agrees with known standard
  - Precision – amount of info/detail in the measure (lack of random variability)
- Validity – extent it measures what you intend
  - Face, content,
  - Criterion-related (infer value on known std): concurrent or predictive (std administered after)
  - Construct validity – do people behave according to a theory that the measure is part of?
  - Ecological – reflects real life

Notes



- Categorical/Discrete
  - Nominal scale – unordered categories
  - Ordinal – have ordering
- Numeric
  - Interval – fixed distance between pts, but arbitrary zero (e.g., celsius temperature)(cannot say X is 200% Y)
  - Ratio – fixed distance, zero=no stuff being measured (e.g., Kelvin temp)

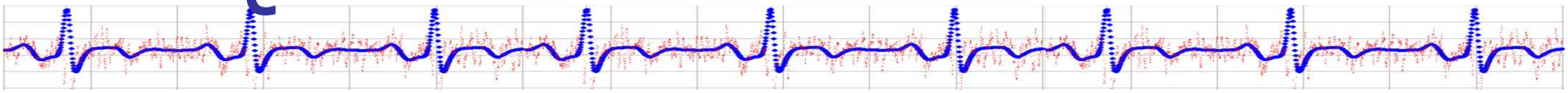
Chapter 9



Using Survey Research

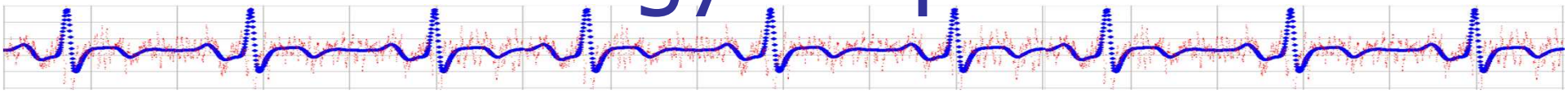
Part I – Questionnaire Design

Questionnaires

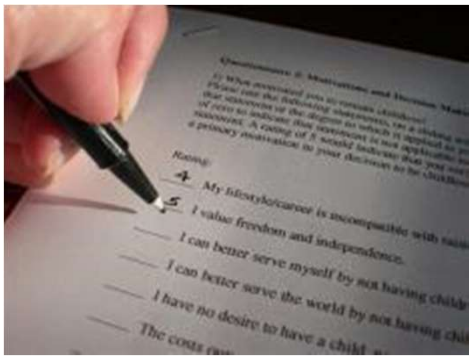


- Asking people to provide responses to questions
- A kind of measure, distinct from the research model it is used in

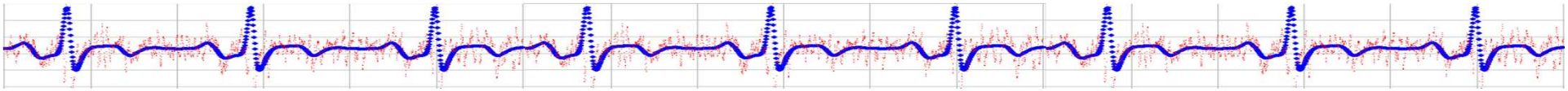
Terminology soup



- Questionnaire = self-report measure = instrument
- Field survey vs. lab instrument/questionnaire
- Composite measure ~ index ~ scale
- Item = question

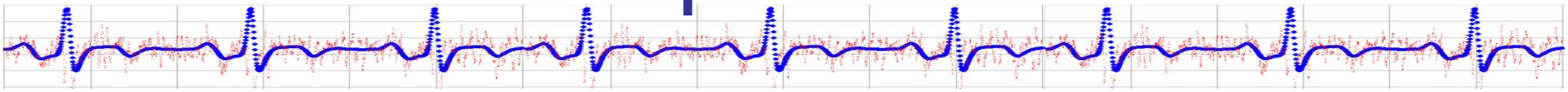


Overview of questionnaire construction



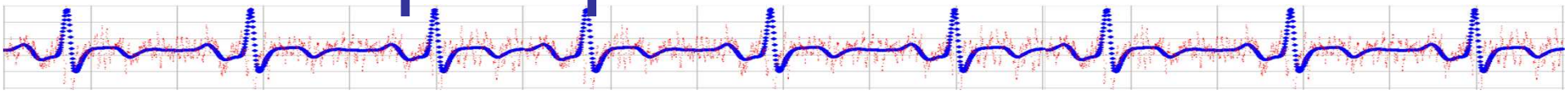
Note: Most of the heuristics on questionnaire design in the text are most appropriate for field surveys

Parts of a questionnaire



- In any study you normally want to collect demographics – usually done through questionnaire
- Single items
- Composite items

Sample questionnaire



Participant ID_____ Date _____

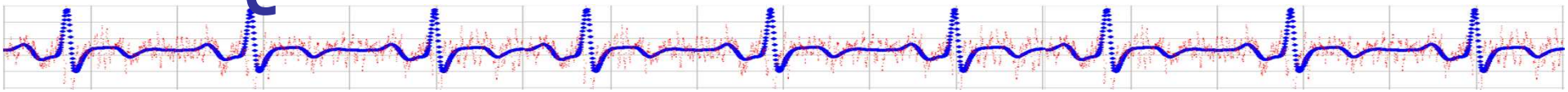
Single item
Single item

Composite measure

Single item

Demographics

Questionnaire construction



Items can be optional; flow often depicted verbally and/or pictorially

14. Have you ever participated in the Model Cities program?

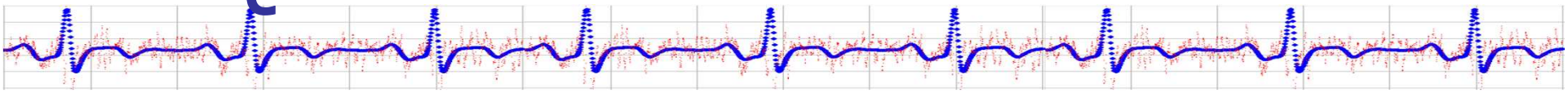
[ ] Yes

[ ] No

If Yes: When did you last attend attend a meeting?

_____

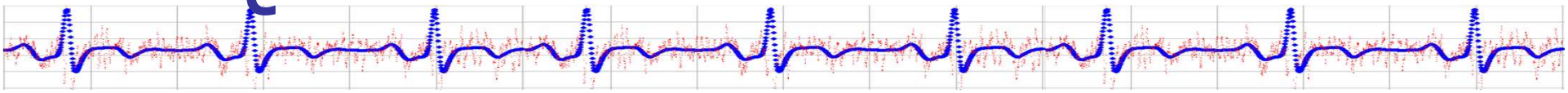
Questionnaire construction



Many heuristics for ordering questions, length of surveys, etc. For example:

- Put interesting questions first
- Demonstrate relevance to what you've told participants
- Group questions in to coherent groups

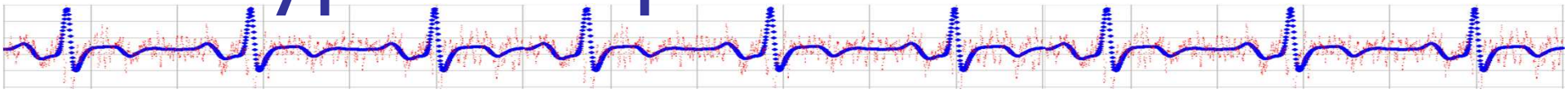
Questionnaire construction



Additional heuristics

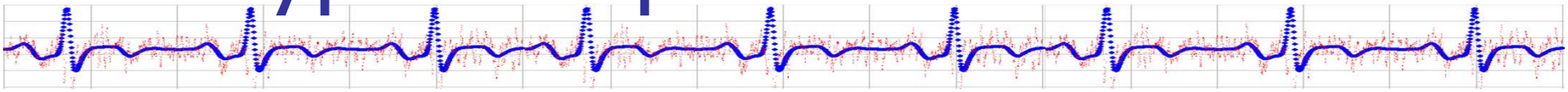
- Organize questions into a coherent, visually pleasing format
- Do not present demographic items first
- Place sensitive or objectionable items after less sensitive/objectionable items
- Establish a logical *navigational path*

Types of questionnaire items



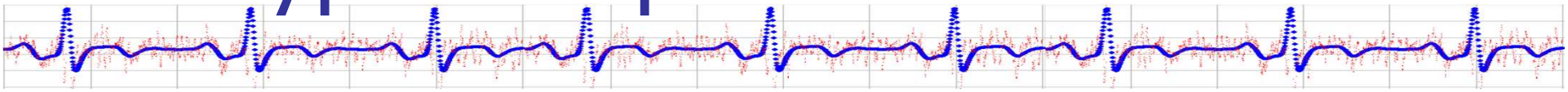
- *Open-ended*
  - Respondents are asked to answer a question in their own words
- *Restricted* (closed-ended)
  - Respondents are given a list of alternatives and check the desired alternative
- *Partially open-ended*
  - An "Other" alternative is added to a restricted item, allowing the respondent to write in an alternative

Types of questionnaire items



- *Rating scale*
  - Respondents circle a number on a scale (e.g., 0 to 10) or check a point on a line that best reflects their opinions
  - Two factors need to be considered
    - Number of points on the scale (5-10)
    - How to label ("anchor") the scale (e.g., endpoints only or each point)

Types of questionnaire items

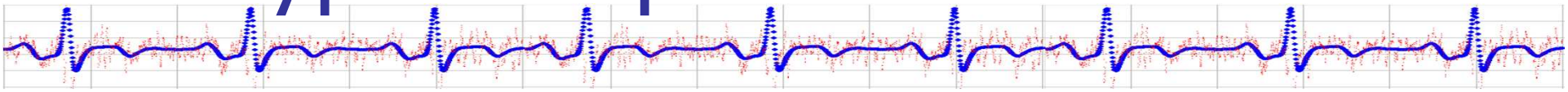


- A *Likert scale*: used to assess attitudes
  - Respondents indicate the degree of agreement or disagreement to a series of statements
    - I am happy.

      Disagree 1 2 3 4 5 6 7 Agree

- A *semantic differential scale*

Types of questionnaire items

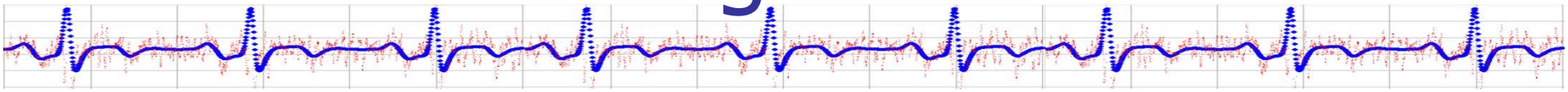


- A *Likert scale*

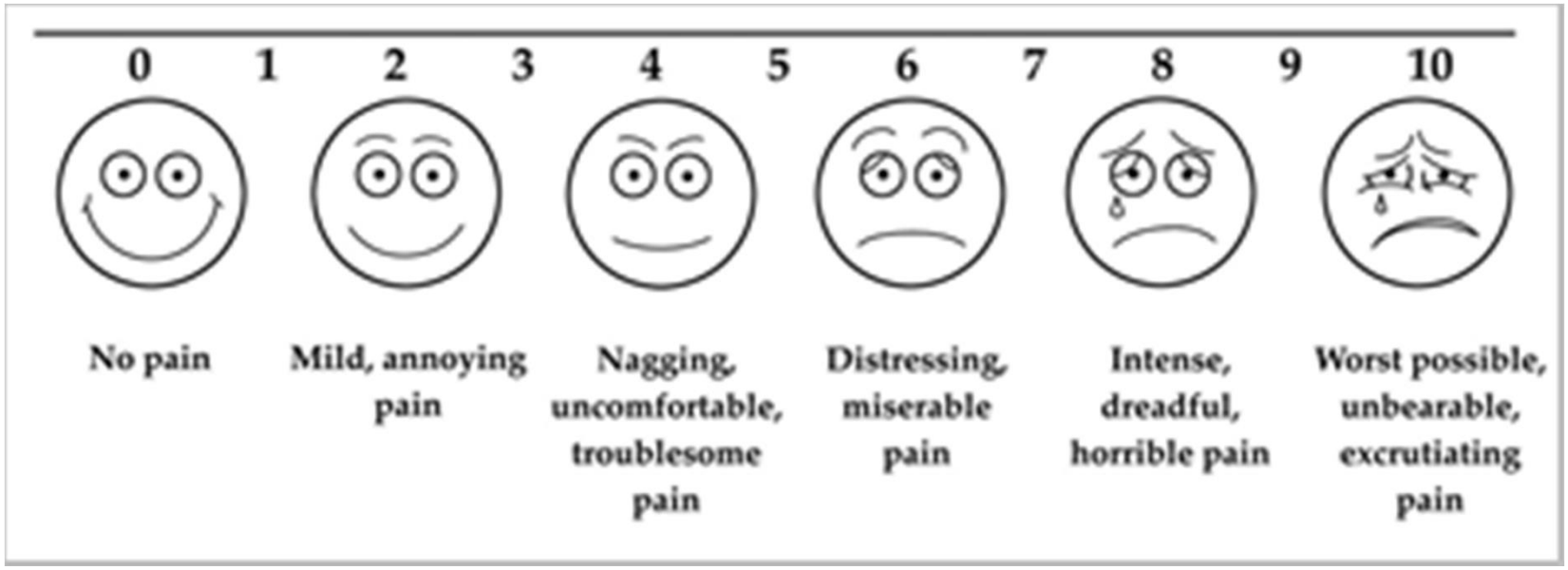
- A *semantic differential scale* allows participants to provide a rating within a bipolar space
  - How are you feeling right now?

    Sad 1 2 3 4 5 6 7 Happy

Visual analog scale

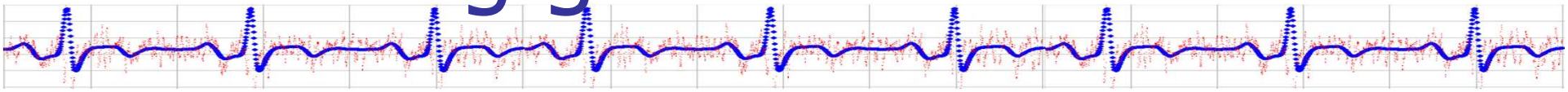


- Indicate position along a line
- A bit more information than quantized scales (e.g., Likert)



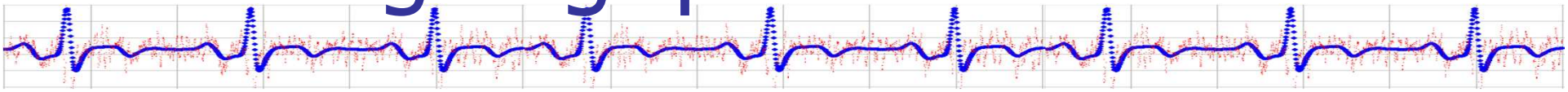
Checklist for homework…

Writing good items



- Use simple words
- Avoid vague questions
- Don’t ask for too much information in one question
- Avoid “check all that apply” items
- Avoid questions that ask for more than one thing
- Soften impact of sensitive questions
- Try to avoid negative statements

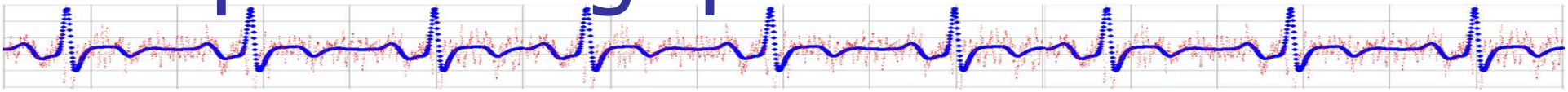
Two most important rules in designing questionnaires?



1. Use an existing validated questionnaire if you can find one.
2. If you must develop your own questionnaire, **pilot test** it and validate it to the extent you can!

Checklist for homework…

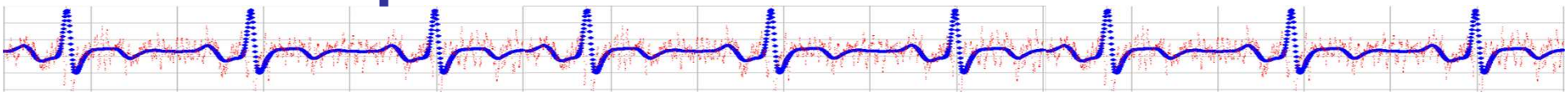
Most important when publishing questionnaire results



- You must either
  - Provide a reference to a previously validated questionnaire, OR
  - Provide the full text of your questionnaire

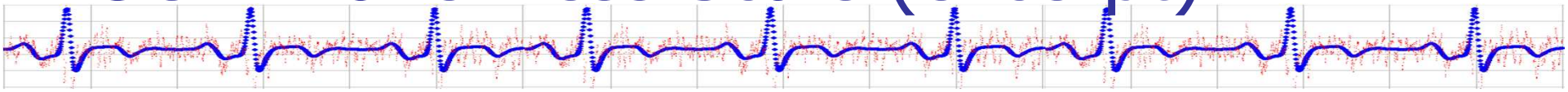
- Without knowing the exact wording and response format (e.g., anchors) readers cannot interpret your results

Composite measures



Classical Test Theory

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

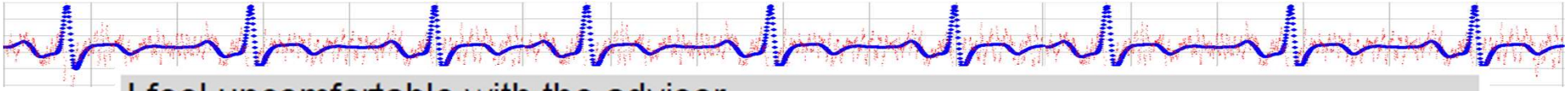
4. I do not feel alone.

NEVER RARELY SOMETIMES ALWAYS

5. I feel part of a group of friends.

NEVER RARELY SOMETIMES ALWAYS

Example composite measure
Working Alliance Inventory (5 of 36 Qs)



I feel uncomfortable with the advisor.

disagree completely	•	•	•	•	•	•	•	agree completely

The advisor and I understand each other.

disagree completely	•	•	•	•	•	•	•	agree completely

I believe the advisor likes me.

disagree completely	•	•	•	•	•	•	•	agree completely

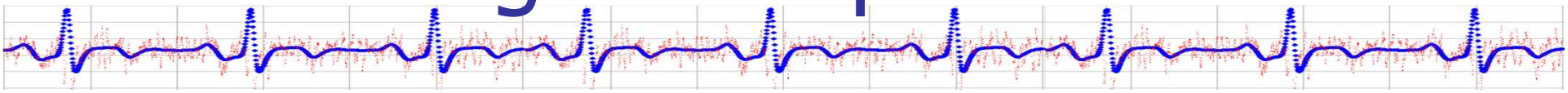
I believe the advisor is genuinely concerned about my welfare.

disagree completely	•	•	•	•	•	•	•	agree completely

The advisor and I respect each other.

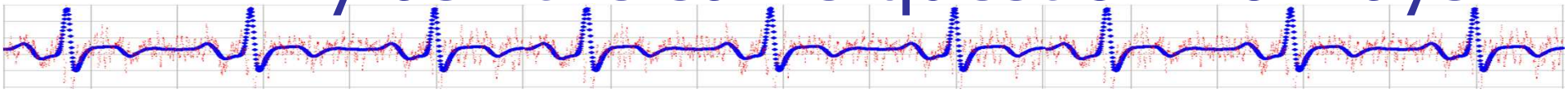
disagree completely	•	•	•	•	•	•	•	agree completely

‘Scoring’ a composite measure



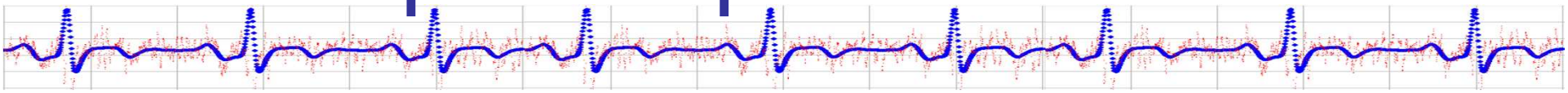
- Generally:
  - Negate negative items
    - Score’ = (max score + 1) – Score
  - Sum scores
- Can normalize by averaging
- Weight items equally unless you have a compelling reason to do otherwise
- Missing data:
  - “Impute the average” by excluding unanswered items from the average

Composite measures:
Why ask the same question 10 ways?



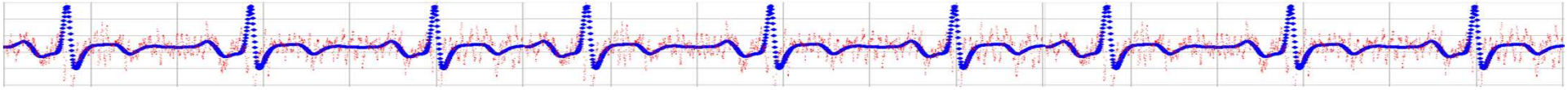
- It is seldom possible to arrive at a single question that adequately represents a complex variable
  - Any single item is likely to misrepresent some respondents (e.g., "church-going")
- A single item may not provide enough variation for your purposes
- Single items give crude assessments; several items may give a more comprehensive and accurate assessment

Composite questionnaire



- Don’t want a numeric measure (meaningless), just want to be able to rank order participants wrt their attitude
- More questions USUALLY provides better reliability
  - Errors in interpretation
  - Errors in association between Q & construct for a given participant
- Why reverse code items? Response bias

Terminology: Factors, subscales & constructs



- Construct
  - A psychological entity that you are interested in measuring (e.g., loneliness, working alliance)
- Factor
  - A construct may have more than one part or dimension or aspect, referred to as "factors" that may be independently assessed by your questionnaire
- Subscale
  - A part of your questionnaire that assesses one factor.
  - Usually: score subscales separately, in addition to aggregate
- Factors can be informed by theory, or emerge from data analysis ("exploratory factor analysis")

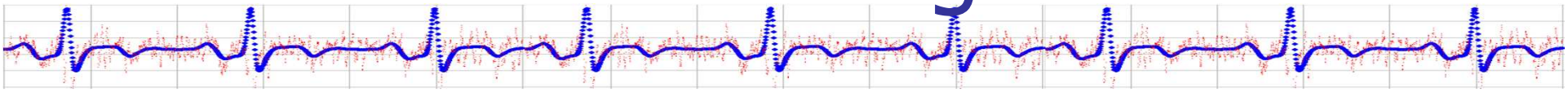
1. (B) I feel uncomfortable with George
2. (T) George and I agree about the things I will need to do to help improve my level of physical activity.
3. (G) I am worried about the outcome of my sessions with George.
4. (T) What I am doing in my discussions with George gives me new ways of looking at physical activity.
5. (B) George and I understand each other.
6. (G) George perceives accurately what my goals are.
7. (B) I find what I am doing with George confusing.
8. (B) I believe George likes me.
9. (G) I wish George and I could clarify the purpose of our sessions.
10. (G) I disagree with George about what I ought to get out of my discussions with him.
11. (T) I believe the time George and I are spending together is not spent efficiently.
12. (G) George does not understand what I am trying to accomplish.
13. (T) I am clear on what my responsibilities are with respect to physical activity.
14. (G) My physical activity goals are important to me.
15. (G) I find what George and I are doing are unrelated to my concerns.
16. (T) I feel that the things I do with George will help me to accomplish the changes that I want.
17. (B) I believe George is genuinely concerned about my welfare.
18. (T) I am clear as to what George wants me to do in our discussions.

T: Task
B: Bond
G: Goal

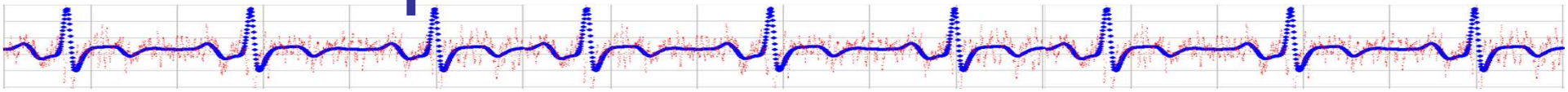
19. George and I respect each other.
20. I feel that George is not totally honest about his feelings toward me.
21. I am confident in George's ability to help me.
22. George and I are working towards mutually agreed upon goals.
23. I feel that George appreciates me.
24. We agree on what is important for me to work on.
25. As a result of my discussions with George I am clearer as to how I might be able to change.
26. George and I trust one another.
27. George and I have different ideas on what my problems are.
28. My relationship with George is very important to me.
29. I have the feeling that if I say or do the wrong things, George will stop working with me.
30. George and I collaborate on setting goals for us to work on.
31. I am frustrated by the things I do with George.
32. We have established a good understanding of the kind of changes that would be good for me.
33. The things that George is asking me to do don't make sense.
34. I don't know what to expect as the result of my discussions with George.
35. I believe the way we are working with my problem is correct.
36. I feel George cares about me even when I do things that he does approve of.

T: Task
B: Bond
G: Goal

“Stress counseling” exercise

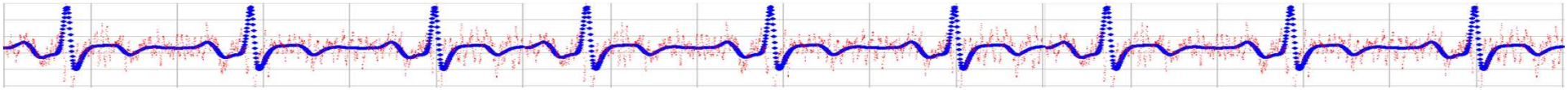


Composite measures

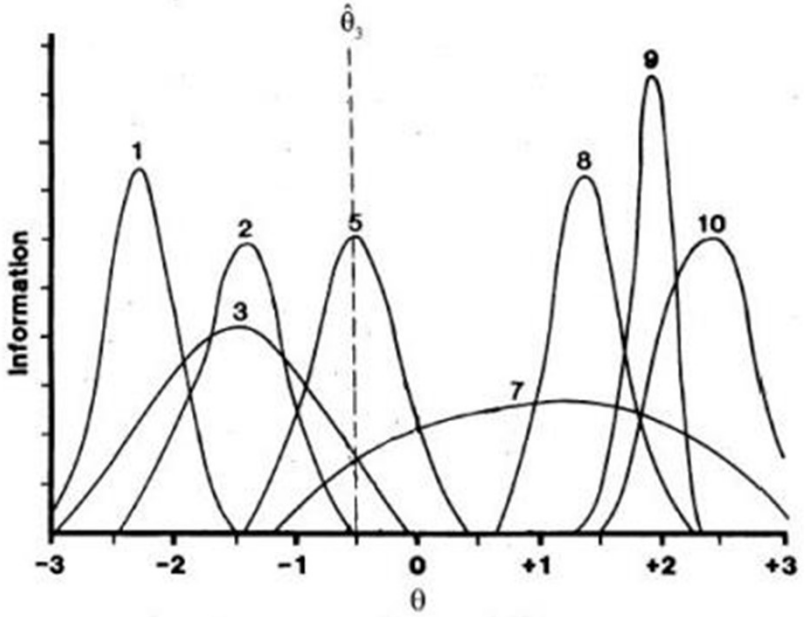


- Indexes (aka "scales") provide an ordinal ranking of respondents with respect to a concept of interest (e.g., liking of computers)
- Usually assessed through a series of related questions.
- Psychological concepts
  - Most have no real meaning, no ultimate definition
  - Ad hoc summaries of experience and observations

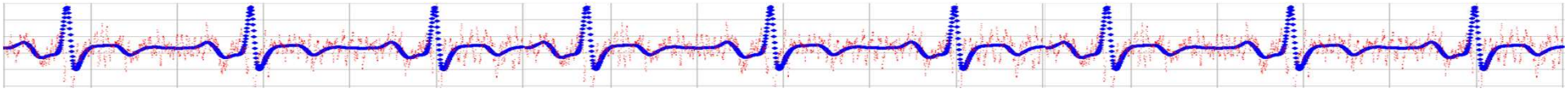
Contrast with IRT-based CAT

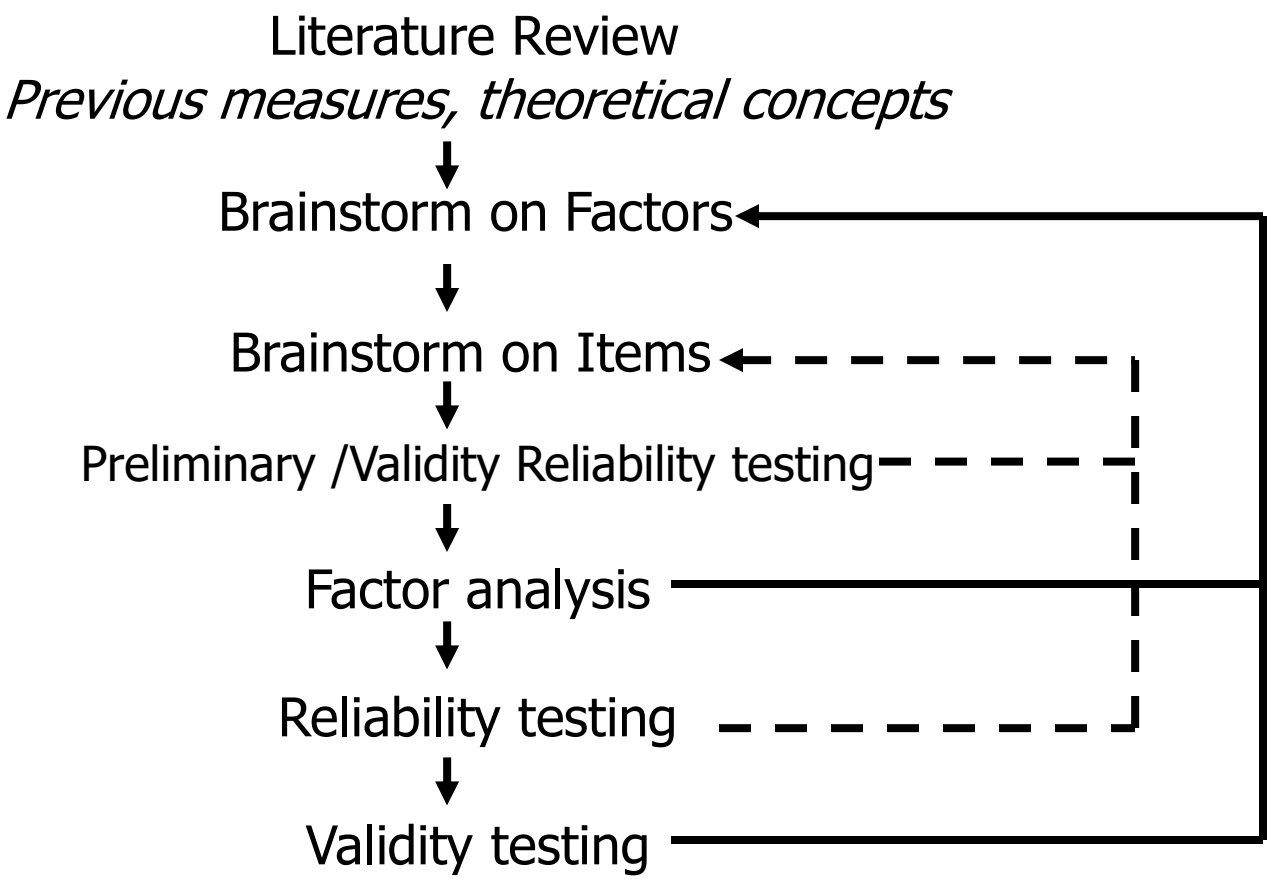


- Item response theory (IRT)
- Computer adaptive test (CAT)

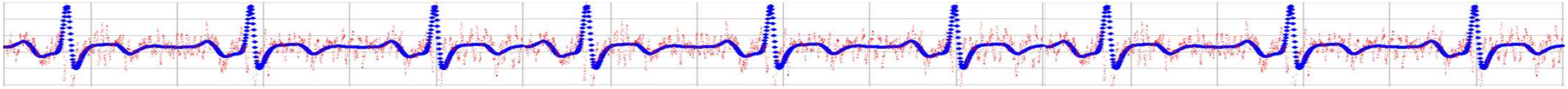


Designing a composite measure



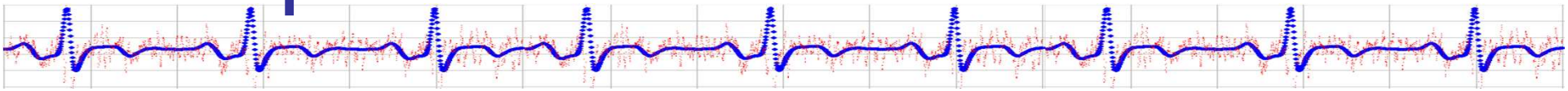


Psychological concepts aka "constructs"



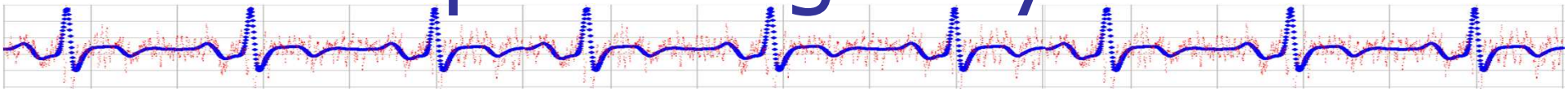
- Concepts are general codifications of experience and observations.
  - Observe differences in social standing -> concept of social status.
  - Observe differences in religious commitment -> concept of religiosity
- Most psychological concepts have no real meaning, no ultimate definitions
- Concepts are ad hoc summaries of experience and observations

Operationalization



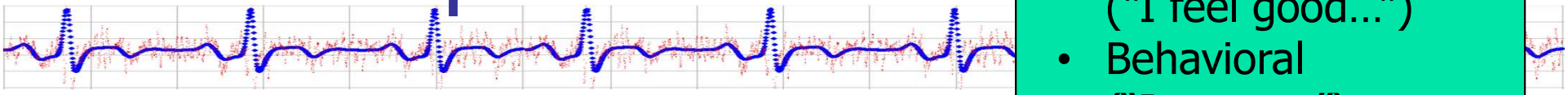
- The process of specifying empirical observations that are indicators of the concept of interest
- Begin by enumerating all the subdimensions ("factors") of the concept
  - Review previous research
  - Use commonsense

Example: Religiosity



- Subdimensions/indicators/factors
  - Ritual involvement
    - E.g., going to church
  - Ideological involvement
    - Acceptance of religious beliefs
  - Intellectual involvement
    - Extent of knowledge about religion
  - Experiential involvement
    - Range of religious experiences
  - Consequential involvement
    - Extent to which religion guides social decisions
- There are others

Example

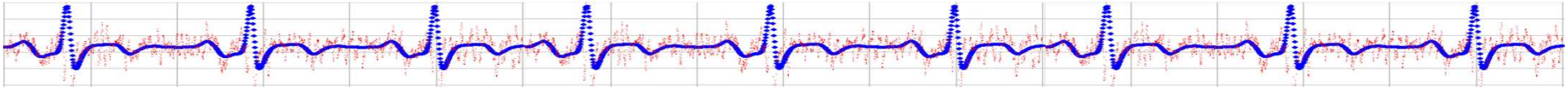


- "NU Husky Fanatic"

1. What are some factors?
2. What are some items per factor?

- Attitudinal ("I like…")
- Emotional ("I feel good…")
- Behavioral ("I go to …")
- Cognitive ("I know a lot about..")

Discriminant indicators

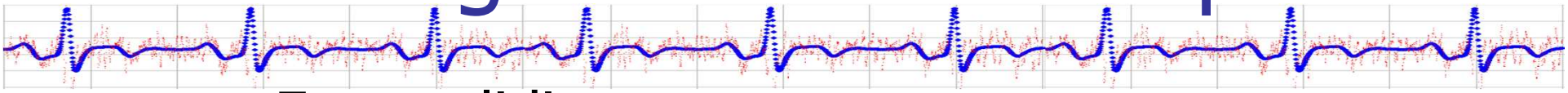


- Also think about related measures which should not be indicators of your construct
- In particular if you will be measuring another related variable, make sure none of your indicators include any attributes of it
- Example

  Want to study the relationship between religiosity and attitudes towards war => including a measure of adherence to "peace on earth" doctrine is not a good idea

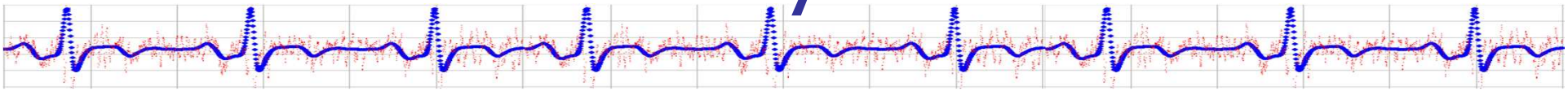
Checklist for homework…

Picking items for a composite



- Face validity
- Unidimensionality
  - All items measure same concept
- Should provide variance in responses
  - Don't pick items that classify everyone one way
  - If you are interested in a binary classification (e.g., liberal vs. conservative), each item should split respondents roughly in half
- Negate up to half of the items to avoid response bias

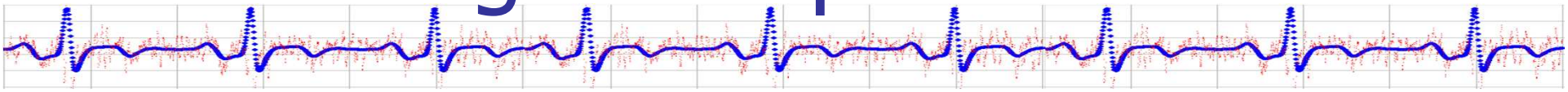
Picking items: Bivariate analysis



Every pair of items should be related, but not too strongly

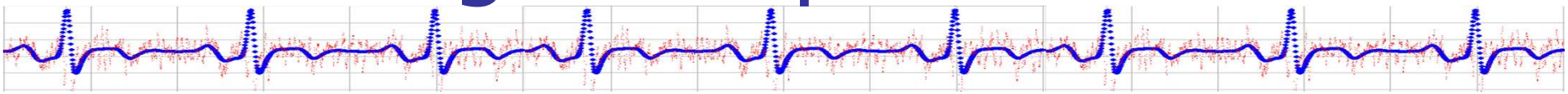
- Scoring high on item A should increase likelihood of scoring high on item B
- But, if two items are perfectly correlated (e.g. one logically implies the other), then one can be dropped

Scoring a composite measure

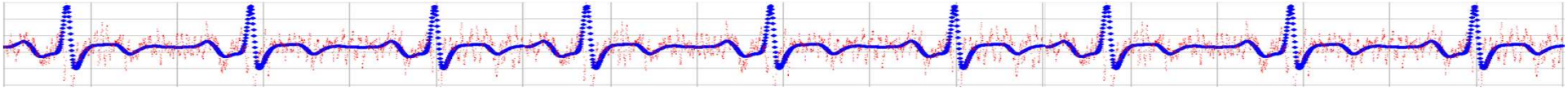


- Average the item scores
- Weight items equally unless you have a compelling reason to do otherwise
- Missing data
  - Omit dataset
  - Impute average/intermediate score
  - "Last value forward" for repeated measures
  - Many other strategies

Validating a composite measure



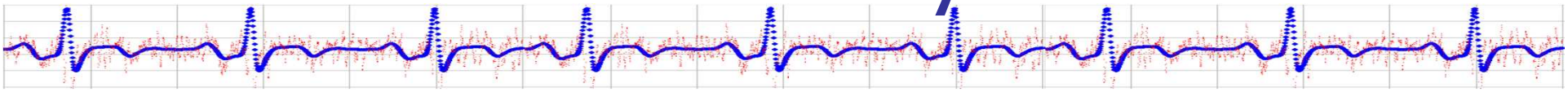
What is a validated measure?



- Has reliability
- Has validity

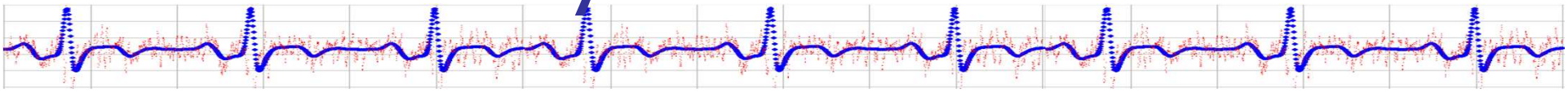
- For psychological measures, these are collectively referred to as a measure's "psychometrics"

Measure reliability



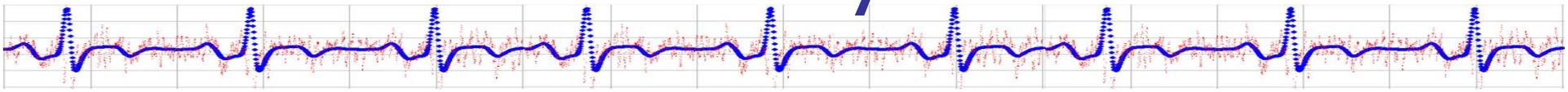
- A reliable measure produces similar results when repeated measurements are made under identical conditions
- Established by:
  - *Test-retest reliability:* Administer the same test twice
  - *Parallel-forms reliability:* Alternate forms of the same test used
  - *Split-half reliability:* Parallel forms are included on one test and later separated for comparison

Reliability



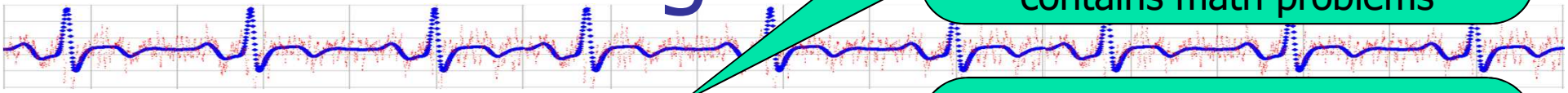
- For composite measure questionnaires, this also encompasses *internal consistency:*
  - Do all of the questions address the same underlying construct of interest?
  - That is, do scores co-vary?
  - A standard measure is Cronbach's alpha
    - 0 = no correlation
    - 1 = scores always covary in the same way
    - 0.7 used as conventional threshold

Measure validity



- A valid measure measures what you intend it to measure
- Validity can be established in a variety of ways…

Establishing vali



- *Face validity:* Subject
  adequacy of content.
  method
- *Content validity:* How
  a test sample beha
  measure?
  - Does each item relate t
  - Do the items collectivel

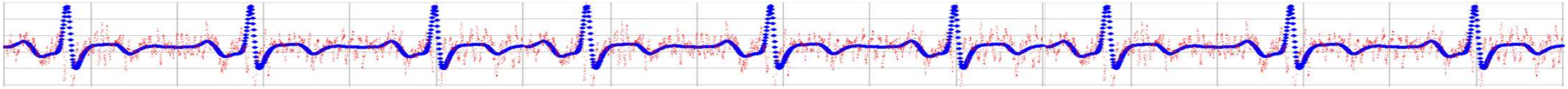
Whether the test "looks valid" to people using/taking it (important sometimes)
E.g., test of math ability contains math problems

Use of recognized domain experts; assess agreement among subject matter expert raters or judges regarding how essential a particular item is

Example: Political attitudes: include items relevant to all the major issues related to such attitudes (e.g., abortion, health care, the economy, and defense)

Example: Final exam covers all material in the course

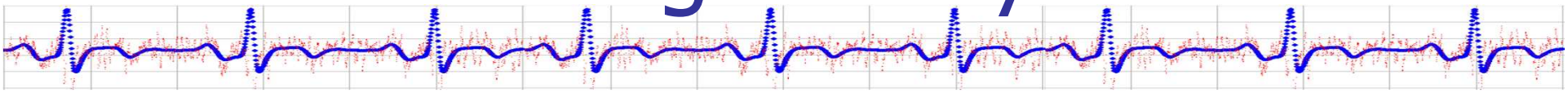
Construct



A variable, not directly observable, that has been developed to explain behavior on the basis of some *theory*

Examples: “intelligence,” “self-esteem,” “achievement motivation”

Establishing validity



- *Construct validity:* Do the results of a test correlate with what is theoretically known about the construct being evaluated? Does it measure what it claims?
  - 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

ng validi

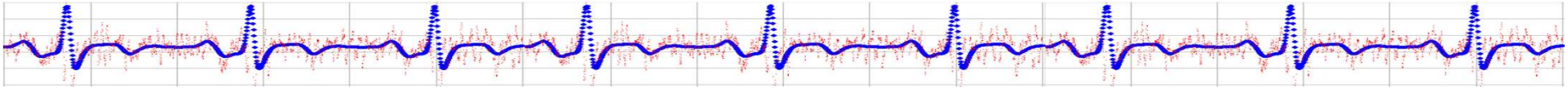
Unlike construct validity, not necessarily related to theory

E.g., Check if survey results correlate with another measure of the same dimension taken at the same time

- *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 **taken at the same time**?
  - Predictive validity: Does test predict behavior at a later time known to be associated with the behavior being measured?

E.g., Survey predicts election results

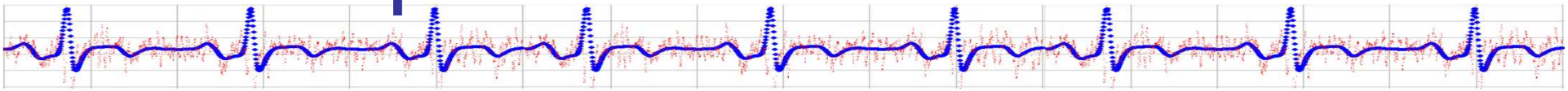
Question from reading assessment



Correlating a questionnaire's result with those from another, established measure is an example of _____.

- Construct validity
- Criterion-related validity
- Face validity
- Kappa validity

Overall process to develop a composite measure



- Identify factors
- Identify items
- Face and content validity for each item
- Check response variance for each item
  (Check floor/ceiling effects)
- Bi-variate analysis
- Test reliability
- Test validity

Good checklist for assignment!