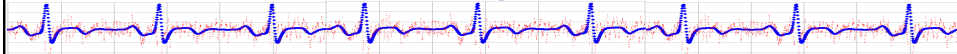


# Empirical Research Methods in Information Science

IS 4800 / CS 6350



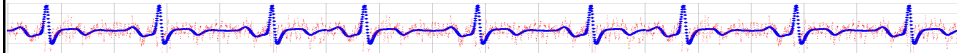
## Lecture 9

Survey Design  
Composite Measure Design  
Survey Administration  
Sampling  
Preview of Hypothesis Testing

1

## Homework Review I3

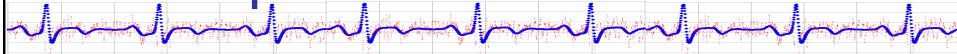
Due now



- Usability/Performance Measures
- Descriptive Stats

3

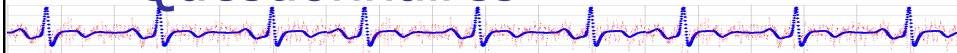
## Chapter 9



### Using Survey Research Part I – Questionnaire Design

4

## Questionnaires

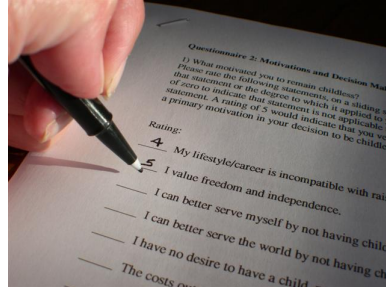


- Asking people to provide responses to questions
- A kind of measure, distinct from the research model it is used in

5

## Terminology Soup

- Questionnaire = Self-Report Measure = Instrument
- Field Survey vs. Lab Instrument/Questionnaire
- Composite Measure ~ Index ~ Scale
- Item = Question



6

## Overview of Questionnaire Construction

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

8

## Parts of a Questionnaire

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

9

## Sample Questionnaire

**Participant ID** \_\_\_\_\_ **Date** \_\_\_\_\_

***Single item***

***Single item***

***Composite measure***

***Single item***

***Demographics***

10

## 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 a meeting?  
\_\_\_\_\_

11

## 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

12

## 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*

13

## 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

14

## 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)

15

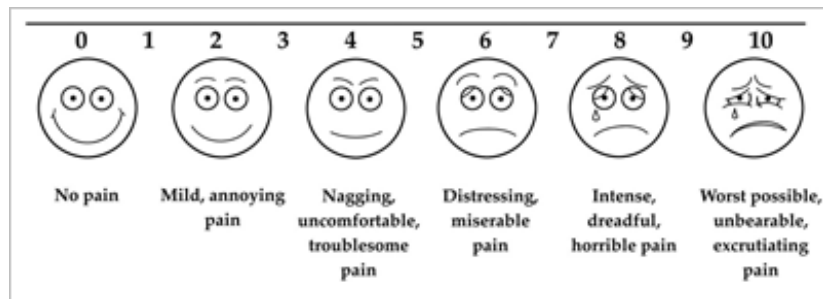
## Types of Questionnaire Items



- A *Likert Scale* is a 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* 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

16

## Visual Analog Scale



17

## 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
- Avoid negative statements (usually)

18



## 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!

19

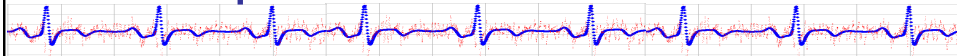
## Most important rules in publishing questionnaire results



- 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

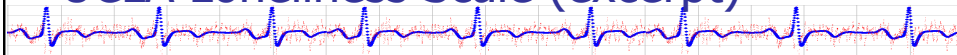
20

# Composite Measures



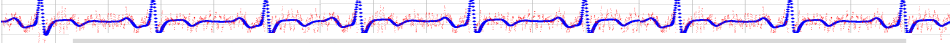
## Classical Test Theory

# Example 'Composite Scale Questionnaire' UCLA Loneliness Scale (excerpt)



<b>1. I feel in tune with the people around me.</b>	NEVER	RARELY	SOMETIMES	ALWAYS
<b>2. I lack companionship.</b>	NEVER	RARELY	SOMETIMES	ALWAYS
<b>3. There is no one I can turn to.</b>	NEVER	RARELY	SOMETIMES	ALWAYS
<b>4. I do not feel alone.</b>	NEVER	RARELY	SOMETIMES	ALWAYS
<b>5. I feel part of a group of friends.</b>	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

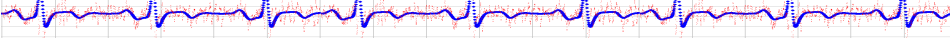
I believe the advisor likes me.  
disagree completely • • • • • • • • agree completely

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

23

## 'Scoring' a Composite Measure

- 
- Generally:
    - Negate negative items
      - $\text{Score}' = (\text{max score} + 1) - \text{Score}$
    - 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
- 24

## 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.

25

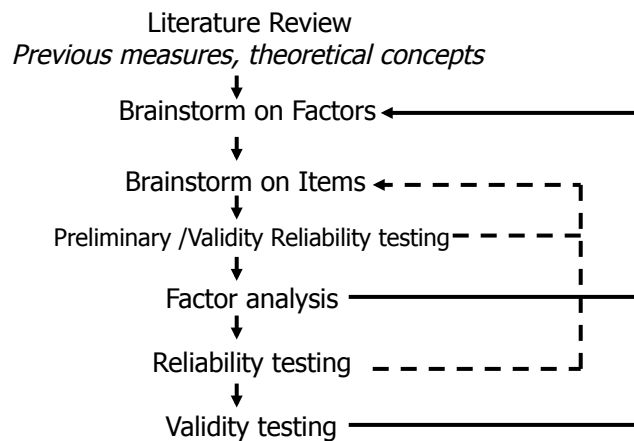
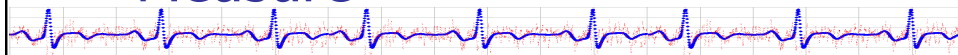
## Terminology: Factors, Subscales & Constructs

- 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")

26

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.

## Designing a Composite Measure



## 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

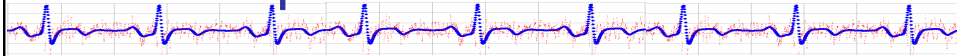
32

## 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 many others)

33

## Example

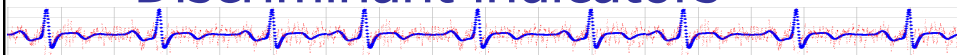


- “NU Husky Fanatic”

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

34

## 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.

35

## 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.

36

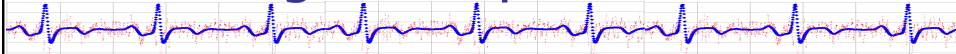
## Picking items: bivariate analysis

- 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.

37

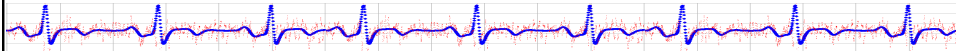


## Validating a Composite Measure



39

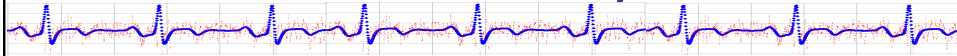
## What is a validated measure?



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

40

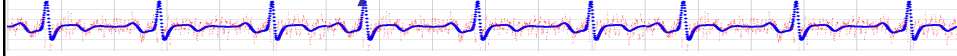
## Measure Reliability



- A reliable measure produces similar results when repeated measurements are made under identical conditions
- Reliability can be established in several ways
  - *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

41

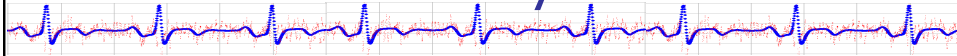
## Reliability



- 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 covary?
  - A standard measure is Cronbach's alpha
    - 0 = no correlation
    - 1 = scores always covary in the same way
    - 0.7 used as conventional threshold

42

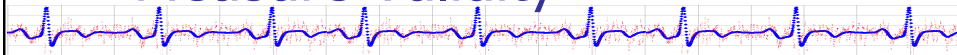
## Measure Validity



- A valid measure measures what you intend it to measure
- 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?
    - Does each item relate to the concept?
    - Do the items collectively cover the concept?

43

## 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

44

## Validation - Summary

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

46

## Overall Process to Develop a Composite Measure

- Identify factors
- Identify items
- Face, content validity for each item
- Check Response Variance for each item
- Bi-variate analysis
- Test reliability
- Test validity

47

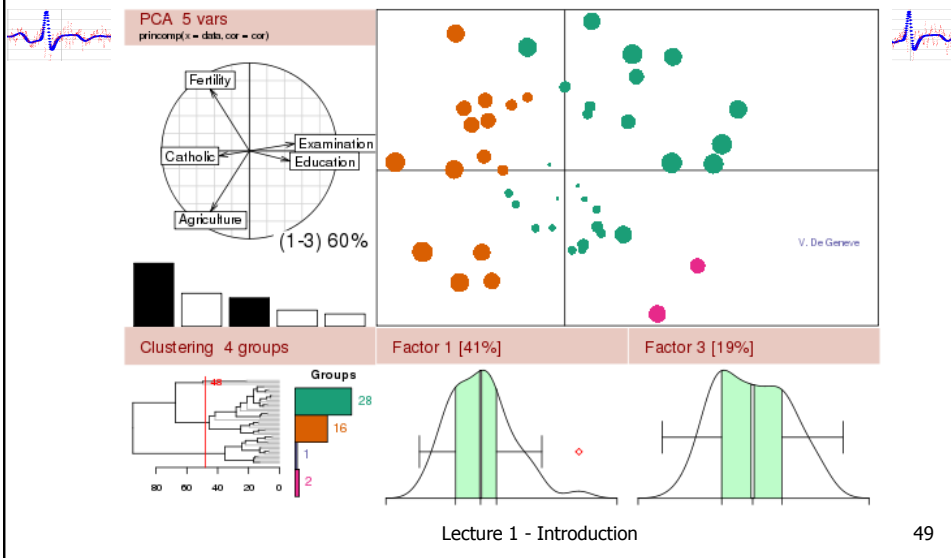
# Developing a New Measure

- 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?

48

# The R Project for Statistical Computing

## Reliability of a Questionnaire



## Cronbach's alpha

negate reverse-coded items first...

```
> install.packages("psych") #one time

> require(psych) #every session
> alpha(data2) #each column of frame = 1 item
Reliability analysis
Call: alpha(x = data2)
raw_alpha std.alpha G6(smc) average_r mean sd
0.11      0.72 0.79 0.39 98 38

... #lots of diagnostic info follows...
```

## Increasing the Reliability of a Composite Questionnaire

- Increase the number of items on your questionnaire
- Standardize the conditions under which the test is administered (e.g., timing procedures, lighting, ventilation, instructions)
- Make sure you score your questionnaire carefully, eliminating scoring errors
- Check to be sure the items on your questionnaire are clearly written and appropriate for those who will complete your questionnaire
- Assess reliability with each item dropped (e.g., "alpha" function in R "psych" package does this).

51

## Exercise – Teams

(to 3:50)

- Design a Composite Measure to assess...X
- Present your survey to the class –
  - Email to Tim or use [Google.com/forms](https://www.google.com/forms)
- be prepared to discuss:
  - Underlying factors considered
  - Bivariate analysis
  - Justification for individual items & format
  - How you would assess reliability
  - How you would assess validity

52

## Questionnaire Administration & Results Analysis

53

## Administering Your Questionnaire



- **MAIL SURVEY**
  - A questionnaire is mailed directly to participants
  - Mail surveys are very convenient
  - Nonresponse bias is a serious problem resulting in an unrepresentative sample
- **INTERNET SURVEY**
  - Survey distributed via e-mail or on a Web site
  - Large samples can be acquired quickly
  - Biased samples are possible because of uneven computer ownership across demographic groups
  - *Check out [surveymonkey.com](http://surveymonkey.com)*

54

## Administering Your Questionnaire



- **TELEPHONE SURVEY**
  - Participants are contacted by telephone and asked questions directly
  - Questions must be asked carefully
  - The plethora of “junk calls” may make participants suspicious
- **GROUP ADMINISTRATION**
  - A questionnaire is distributed to a group of participants at once (e.g., a class)
  - Completed by participants at the same time
  - Ensuring anonymity may be a problem

55



## Administering Your Questionnaire



- *INTERVIEW*
  - Participants are asked questions in a face-to-face structured or unstructured format
  - Characteristics or behavior of the interviewer may affect the participants' responses

56

## Mechanical Turk



- Amazon mechanical turk is a Crowdsourcing tool developed by Amazon.
- Used to have people perform small tasks for micropayments.
- Developed by Peter Cohen at Amazon for its internal use. [To find duplicates among webpages in 2005]
- Amazon recommends paying \$6/hour. But, people have reported some pay as low as \$1/hour (\$2/hr typical)

# Mechanical Turk

amazonmechanicalturk Artificial Intelligence

Your Account | HITs | Qualifications

Introduction | Dashboard | Status | Account Settings

**Mechanical Turk is a marketplace for work.**  
 We give businesses and developers access to an on-demand, scalable workforce.  
 Workers select from thousands of tasks and work whenever it's convenient.  
**334,267 HITs** available. [View them now.](#)

### Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](#)

**As a Mechanical Turk Worker you:**

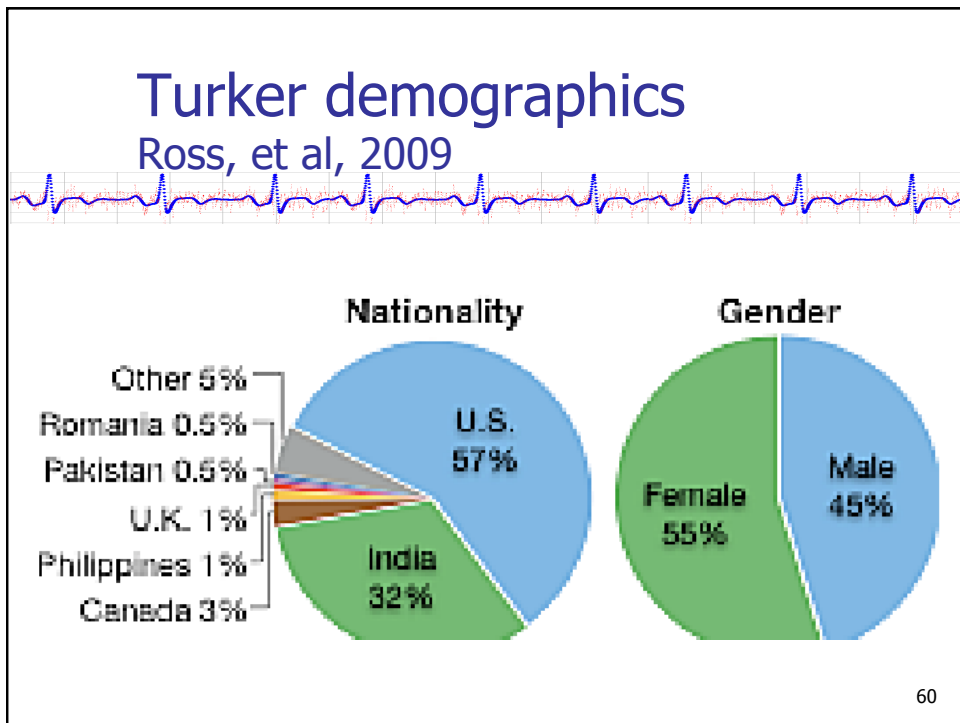
- Can work from home
- Choose your own work hours
- Get paid for doing good work

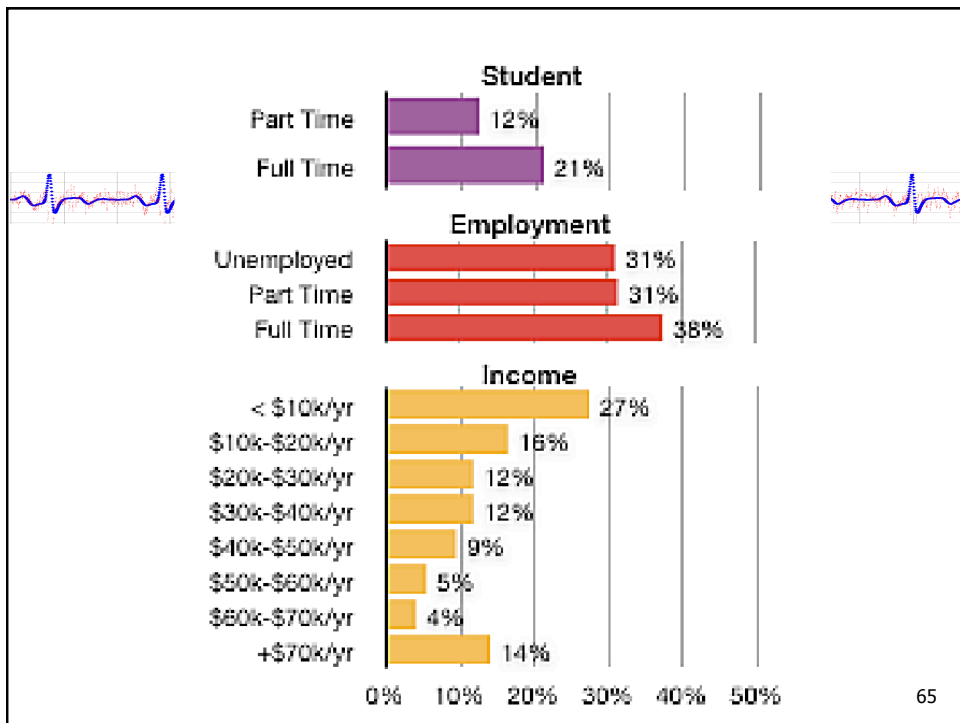
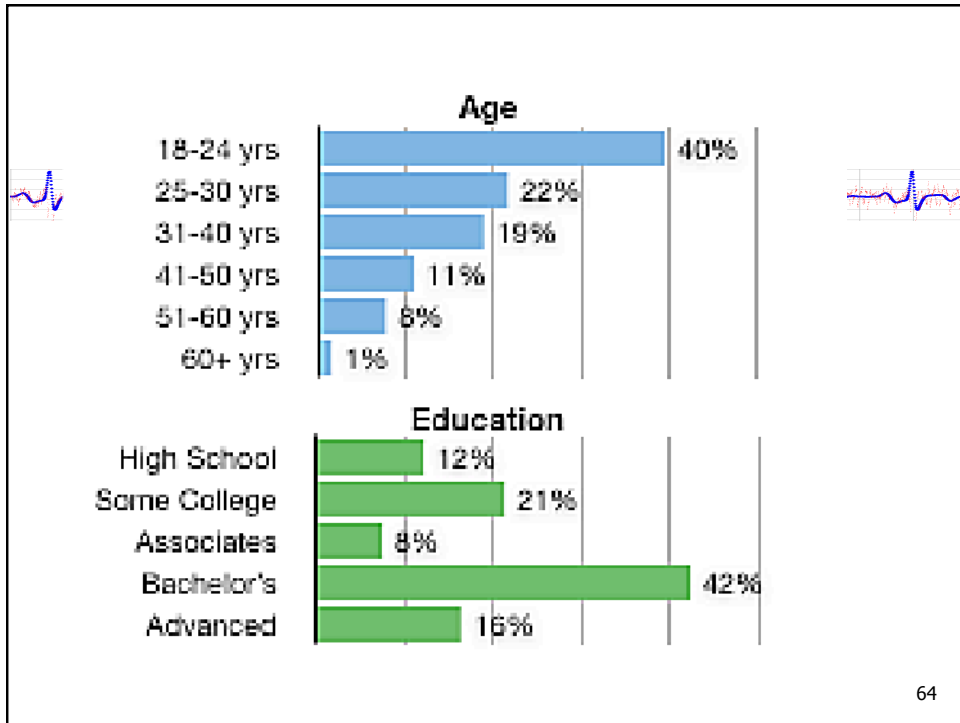
### Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Register Now.](#)

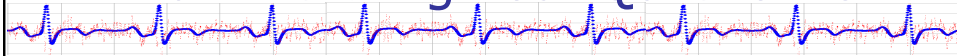
**As a Mechanical Turk Requester you:**

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results





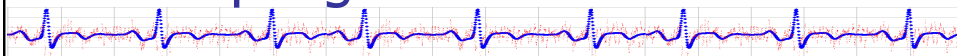
## Administering Your Questionnaire



- In general
  - Personal techniques (interview, phone) provide higher response rates, but are more expensive and may suffer from bias problems.

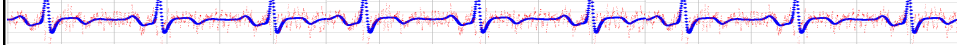
66

## Sampling



67

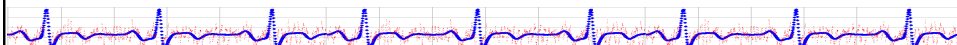
## Sampling



- Sometimes you really can measure the entire population (e.g., workgroup, company), but this is rare...
- “Convenience sample”
  - Cases are selected only on the basis of feasibility or ease of data collection.

68

## Acquiring A Survey Sample



- You should obtain a *representative sample*
  - The sample closely matches the characteristics of the population
- A *biased sample* occurs when your sample characteristics don't match population characteristics
  - Biased samples often produce misleading or inaccurate results
  - Usually stem from inadequate sampling procedures

69

## Sampling Techniques



### ■ *Simple Random Sampling*

- Randomly select a sample from the population
- *Random digit dialing* is a variant used with telephone surveys
- Reduces systematic bias, but does not guarantee a representative sample
  - Some segments of the population may be over- or underrepresented

70

## Sampling Techniques



### ■ *Systematic Sampling*

- Every  $k^{\text{th}}$  element is sampled after a randomly selected starting point
  - Sample every fifth name in the telephone book after a random page and starting point selected, for example
- Empirically equivalent to random sampling (usually)
  - May still result in a non-representative sample
- Easier than random sampling

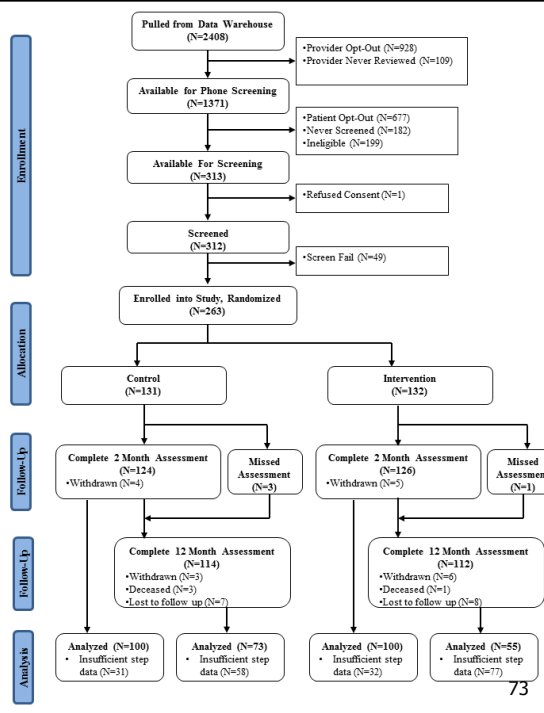
71

# Sampling Techniques

- **Stratified Sampling**
  - Used to obtain a representative sample
  - Population is divided into (demographic) strata
    - Focus also on variables that are related to other variables of interest in your study (e.g., relationship between age and computer literacy)
  - **A random sample of a fixed size is drawn from each stratum**
  - May still lead to over- or underrepresentation of certain segments of the population
- **Proportionate Sampling**
  - Same as stratified sampling except that the proportions of different groups in the population are reflected in the samples from the strata

72

## Example Stratified Sampling



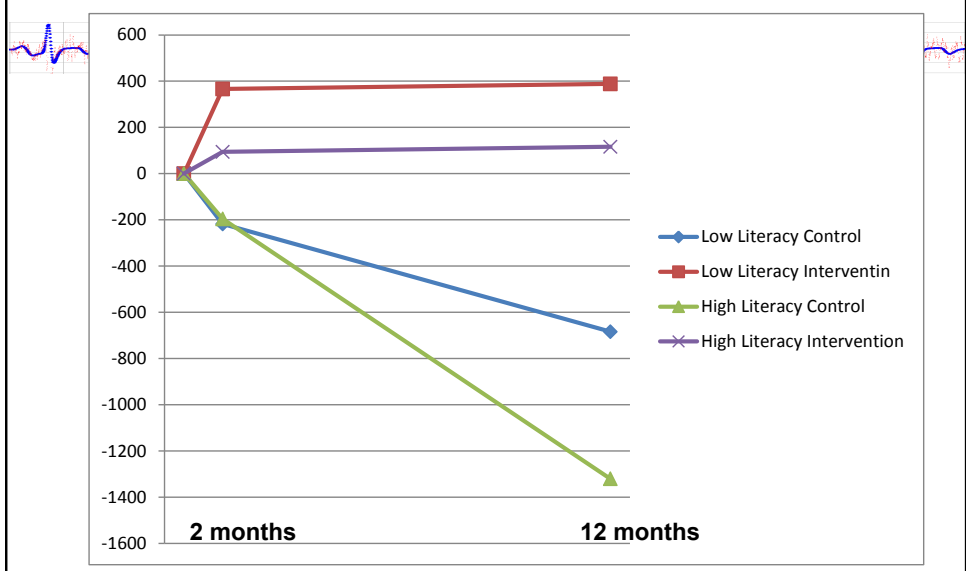
73

# Low Literacy Geriatrics Pts

NIA R01, N=263, 55+

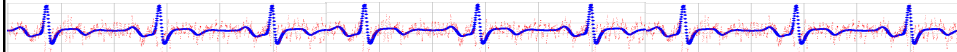


# Low Literacy Geriatrics Pts





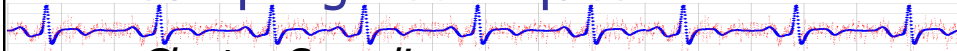
## Sampling Example:



- You want to conduct a survey of job satisfaction of all employees but can only afford to contact 100 of them.
- Personnel breakdown:
  - 50% Engineering
  - 25% Sales & Marketing
  - 15% Admin
  - 10% Management
- Examples of
  - Stratified sampling?
  - Proportionate sampling?

76

## Sampling Techniques



- *Cluster Sampling*
  - Used when populations are very large
  - The unit of sampling is a group (e.g., a class in a school) rather than individuals
  - Groups are randomly sampled from the population (e.g., ten classes from a particular school)

77

## Sampling Techniques

### ■ *Multistage Sampling*

- Variant of cluster sampling
- First, identify large clusters (e.g., school districts) and randomly sample from that population
- Second, sample individuals from randomly selected clusters
- Can be used along with stratified sampling to ensure a representative sample
- Note: Multilevel, hierarchical statistical analysis can tease apart differences due to individual vs. cluster

78

## Sampling

- Most statistics assume a random sample.

79

## Sample size

- In all empirical research, you should motivate your *sample size*
  
- B&A Ch 9 provide formula for estimating sample size for binomial descriptive studies.
  - For binomial (two category) measures
  - Based on
    - Amount of acceptable *sampling error*
    - Expected magnitude of population proportions

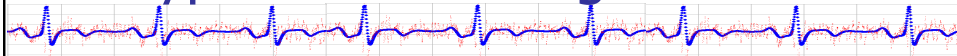
80

## Homework

- Do Homework I4 – Designing a Composite Self-Report Measure
  - Design a new composite self-report measure to assess a person's "homework procrastination". Assume it only has one factor, but use at least five scale items. Incorporate information from at least one literature reference. Assess the face and content validity of your measure and work through a bivariate analysis of your items.
  - Implement questionnaire on surveymonkey.com, Google forms, or similar
  - 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. Email your questionnaire to the class (is4800-all@ccs.neu.edu). (You are also obligated to reply to any questionnaires mailed to you within 48 hrs.)
  - Compute the reliability (internal consistency) of your measure using R Alpha. Compute descriptive statistics for your measure and any other items you may have included on the questionnaire. 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 2/16.

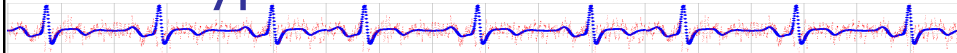
82

## Hypothesis Testing Preview



83

## A typical scenario



- Between-subject design
- Let every subject try both Wizziword & Word
- Measure performance
- Research Hypothesis:
  - Wizziword is better

85

## What if ...

- You can test every subject in your population, AND
- There is no measurement error?
- Nothing else that could cause "error variance"
- 
- Compute descriptives for two treatments
- If Wizziword perf > Word perf conclude H1 is supported
- No uncertainty, No 'p' value

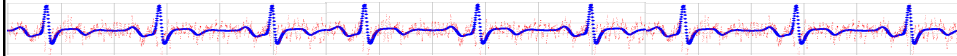
86

## Basic Process of Hypothesis Testing

- H1: Research Hypothesis:
  - Population 1 is different than Population 2
- H0: Null Hypothesis:
  - No difference between Pop 1 and Pop 2
  - *The difference is "null"*
- Compute  $p(\text{observed difference}|H0)$ 
  - 'p' = probability observed difference is due to random variation
- If  $p < \text{threshold}$  then reject H0 => accept H1
  - p typically set to 0.05 for most work
  - p is called the "level of significance"

87

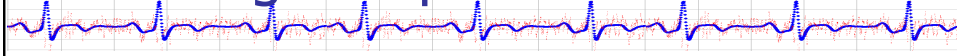
## Other ways of thinking about this...



- “Innocent until proven guilty.”
- How surprising would this result be if there really were no difference?
- Why do things this way???

88

## The grand plan



- $\chi^2$  tests
  - For nominal measures
  - Can apply to a single measure
- Correlation tests
  - For two numeric measures
- t-test for independent means
  - For categorical IV, numeric DV

89