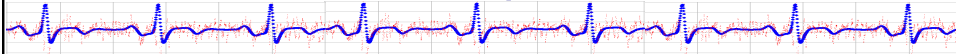


Empirical Research Methods in Information Science

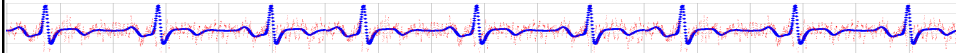
IS 4800 / CS6350



Lecture 8 Miscellaneous Measures

1

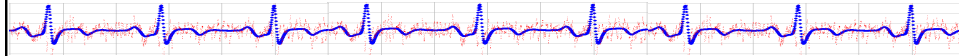
Homework I3 - Due Tuesday Issues?



- Conduct a small usability study
 - Descriptive
 - Quantitative
 - At least Two tasks
 - At least Two measures
 - At least Three subjects

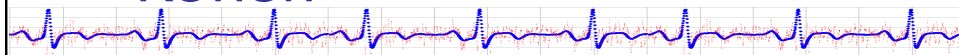
3

Review



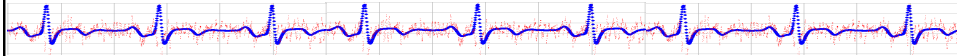
	Number of Variables	Number of IV Levels	Manipulation
Descriptive	1	NA	NA
Demonstration	$\geq 1?$	1	✓
Correlational	≥ 2	NA	NA
Experimental	≥ 2	≥ 2	✓

Questionnaire Validation - Review



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

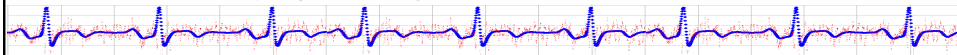
Scales of Measurement - Review



- Nominal
- Ordinal
- Interval
- Ratio

6

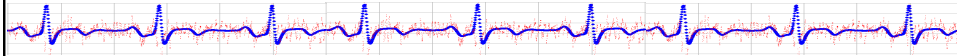
Measures of Center: Decision Rule



- Nominal
 - Mode
- Ordinal
 - Median
- Interval, Ration & Normal & No Outliers
 - Mean
- Else
 - Median

7

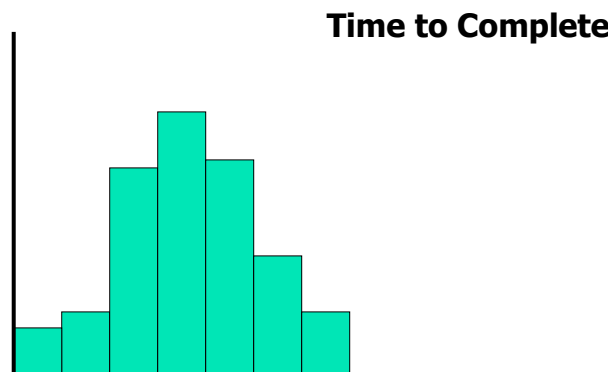
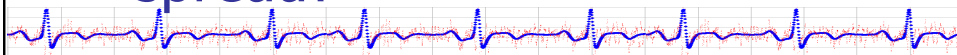
Measures of Spread: Decision Rule



- Nominal, Ordinal
 - no measure of spread
- Interval, Ratio & Normal & no outliers
 - SD
- Else:
 - IQR

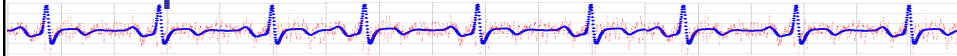
8

Which measures of center and spread?

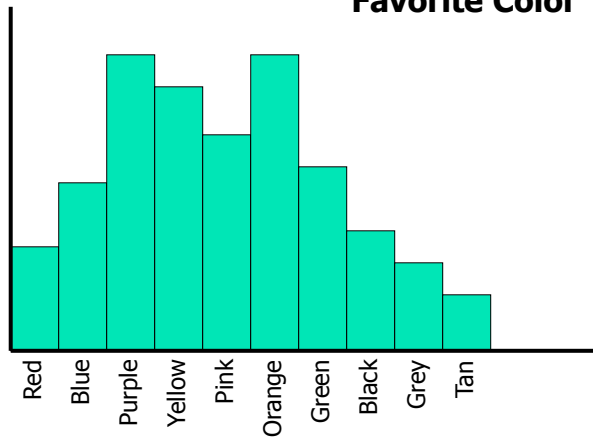


9

Which measures of center and spread?

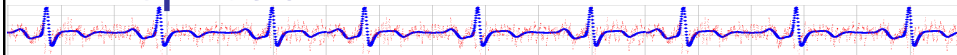


Favorite Color

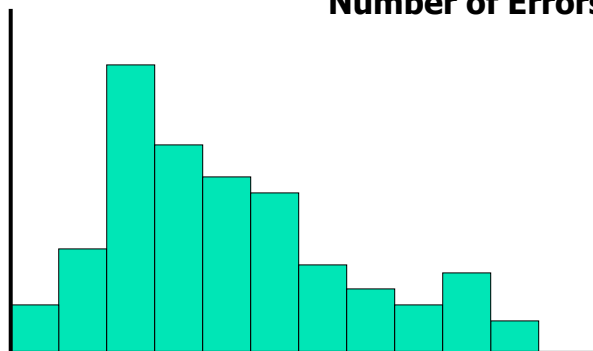


10

Which measures of center and spread?

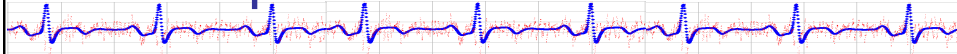


Number of Errors



11

Chapter 8



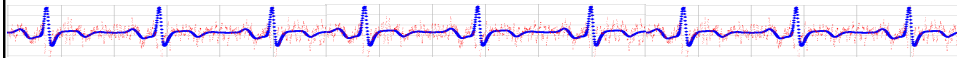
Using Nonexperimental Research
Observational
Miscellaneous designs
Meta-analyses

**Observational
Research**



**Nonexperimental
Research**

Example: Handheld ECAs



- Research Question:
 - Do people exhibit the same nonverbal conversational behavior when talking to a 2" tall character than when talking to another person face-to-face?
- Exercise:
 - Design the study



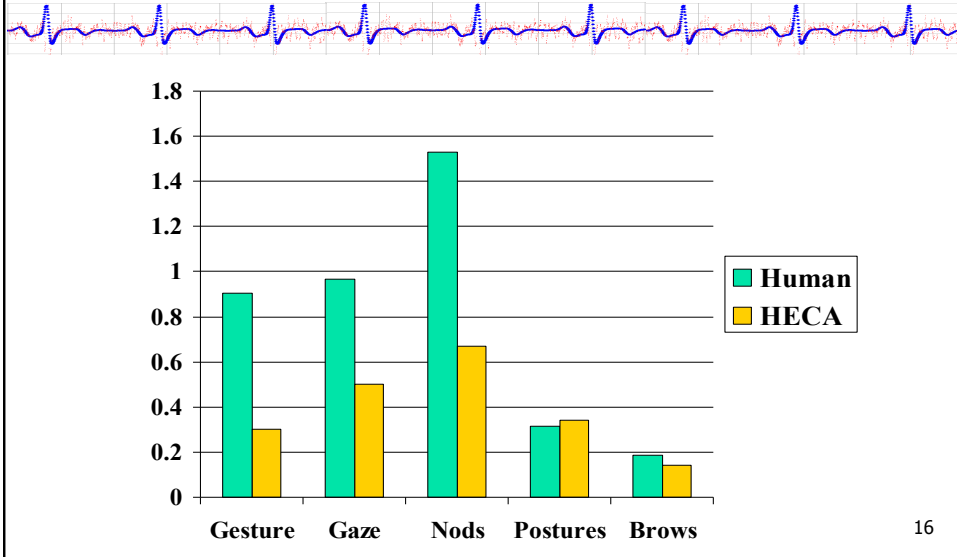
14

Example: Handheld ECAs



15

Results



16

Observational Research aka Behavior Coding

Watching people and quantifying
their behavior

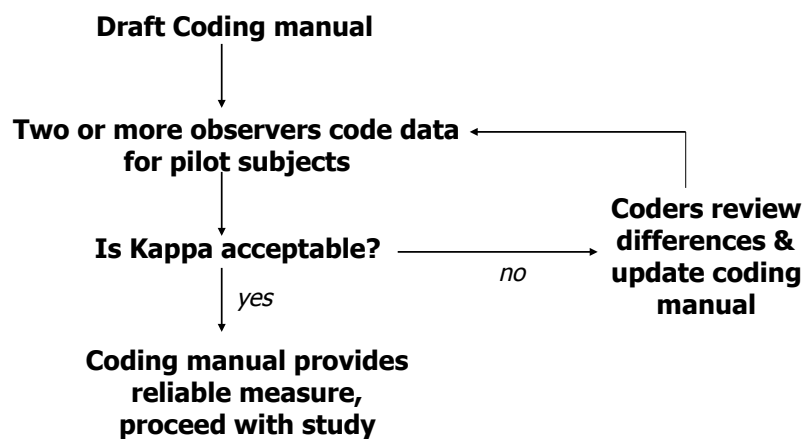
17

Defining Behavioral Categories

- Only need enough detail to provide a reliable measure.
 - What is reliability?
 - How to measure it?
- e.g. do people in the student center get more rude 10 minutes before class times?

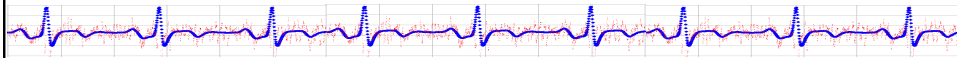


Defining a behavioral protocol



19

Developing Behavioral Categories



- Categories must be operationally defined
- Behavioral categories must be clearly defined to avoid ambiguity

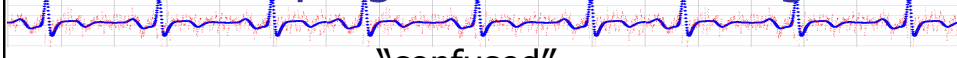
- “flailing arms around”

vs.

- “moved arms from below to above waist and back more than 3 times per minute”

20

Developing Behavioral Categories



“confused”

vs.

(“clicked mouse at least 5 times on inappropriate menu”

OR

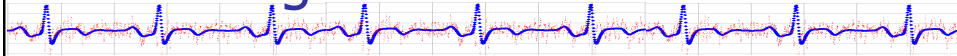
“gazed at interface with mouth open AND no mouse clicks or keyboard presses for 5 minutes”)

AND

“furrowed brows”

21

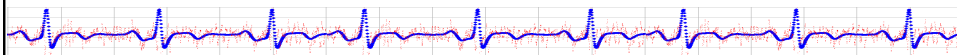
Coding Manual



- You should write your behavior identification rules down so that you could give them to someone else to follow reliably.
- You should also write down the sampling and coding methods you will use, as well as your recording instrument (e.g., paper form).

22

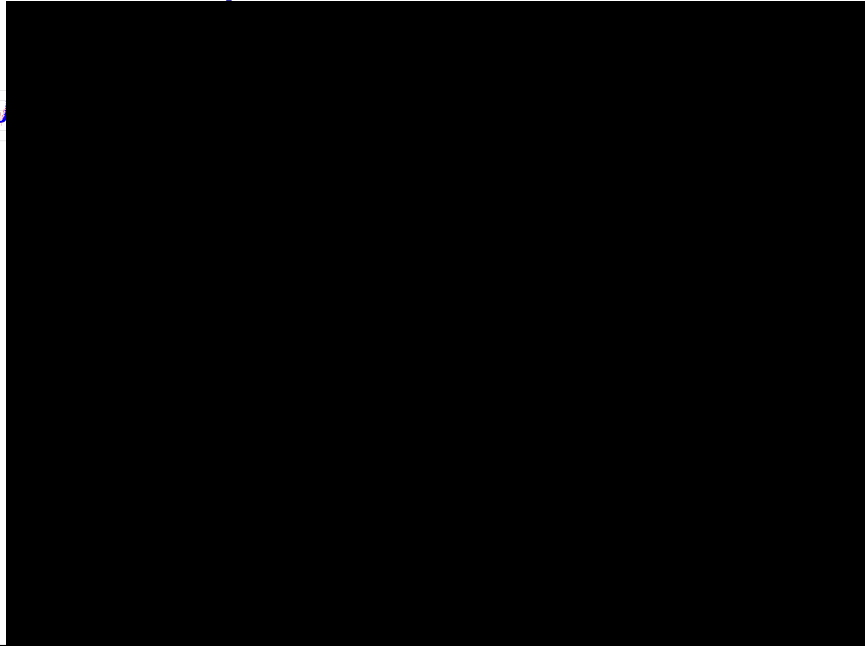
Quantifying Behavior: What is the metric?



- *Frequency Method*
 - Record the frequency with which a behavior occurs within a time period
- *Duration Method*
 - Record how long a behavior lasts
- *Intervals Method*
 - Divide the observation period into several discrete time intervals (e.g., ten 2-minute intervals), and record whether a behavior occurs within each interval

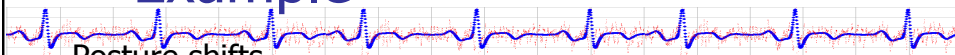
24

Example: Code Posture Shifts



25

Example



■ Posture shifts

- Body part
 - Upper body
 - Lower body
 - Both
- Type
 - Shift
 - Return
- Energy level
 - 0-100%
- Hand gestures and other communicative behavior does not count – nor their effects.
- Video reviewed and start/stop/type coded.
- From this, we can compute frequency, duration, or intervals

StartTime	EndTime	BodyPart	Type	Energy
00:00:03	00:00:04	Upper	Return	50%
...

26

Posture Shifts

Duration, Frequency, or Interval Measures?

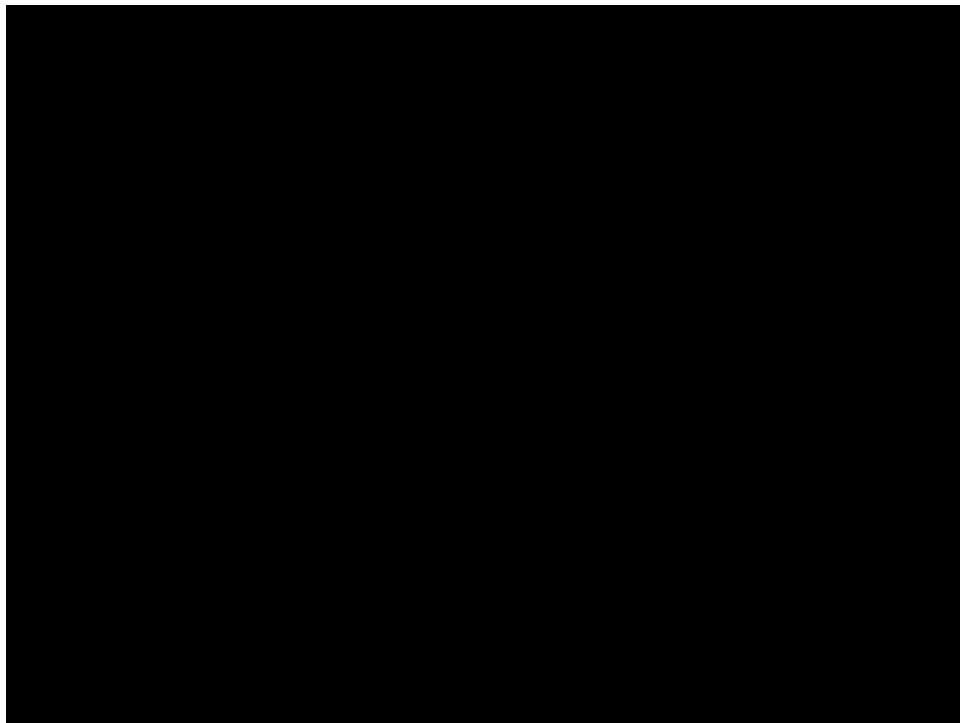


Posture shifts with respect to discourse segment

	Monologues (0.06/s)			Dialogues (0.07/s)		
	ps/s	ps/int	energy	ps/s	ps/int	energy
Inter-dseg	<u>0.340</u>	0.837	0.832	<u>0.332</u>	0.533	0.844
intra-dseg	<u>0.039</u>		0.701	<u>0.053</u>		0.723

Lecture 1 - Introduction

27



Tools for Coding: ANVIL



Coping With Complexity in Observational Research

- **Recording**
 - Use a recording device to make a record of behavior for later review
- **Time Sampling**
 - Scan subjects for a specific period (e.g., 30 seconds), and then record your observations during the next period
- **Individual Sampling**
 - Select a subject and observe behavior for a given period (e.g., 30 seconds), and then shift to another subject and repeat observations

31

Coping With Complexity in Observational Research

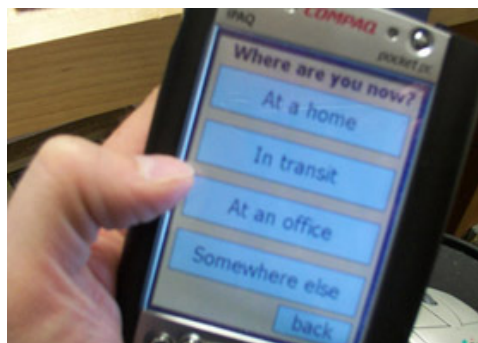
- *Event Sampling*

- Select one behavior for observation and record all instances of that behavior
- It is best if one behavior can be specified as more important than others

32

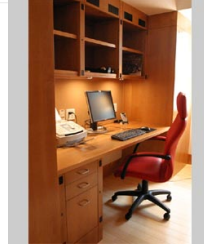
Coping With Complexity in Observational Research

- Ecological momentary assessment
- Intelligent/Context Aware EMA
- What kind of sampling is this?



33

Smart Rooms – e.g. PlaceLab



34

Issues in Observational Research

- IRB issues with video/audio recording?
- Behavior vs. Function/Intent

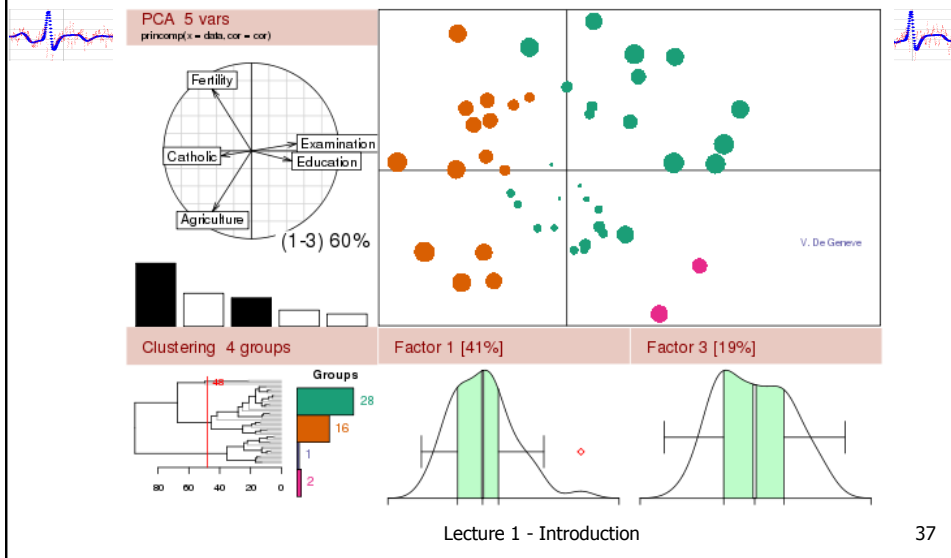
35

Evaluating Interrater Reliability

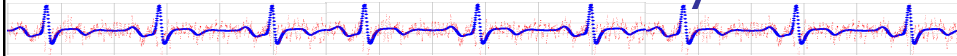
- You must establish reliability of observations from multiple observers (*interrater reliability*)
- Most common/acceptable method for evaluating interrater reliability for a nominal measure, 2 raters
 - *Cohen's Kappa*
 - Allows you to determine if agreement observed is due to chance
 - Kappa of 0.70 or more indicates acceptable interrater reliability

36

The R Project for Statistical Computing Interrater Reliability



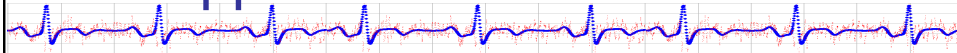
Example R data setup for interrater reliability



Time	Judge1	Judge2
1	together	together
2	apart	apart
3	together	together
4	apart	together
5	apart	apart
6	together	together
7	together	together

45

Kappa



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

> require(psych) #every session

> wkappa(table(data$Judge1,data$Judge2))

$kappa [1] 0.6

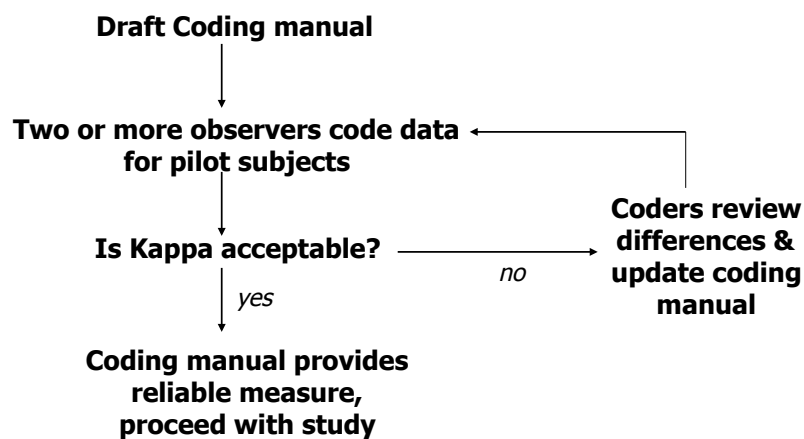
$weighted.kappa [1] 0.2
#accounts for distance of each discrepancy
#=# how bad different disagreements are
#ignore for now
```

Other statistics for inter-rater reliability

- Fleiss' kappa
 - Nominal, >2 raters
- Kendall's τ , or Spearman's rho
 - Ordinal, 2 raters (not testing absolute match)
- Pearson correlation coefficient
 - Interval or ratio, 2 raters (not testing absolute match – only whether linearly related)
- Intraclass correlation coefficient
 - Interval or ratio, 2+ raters
 - See 'icc' function in 'irr' R package.
- See Hallgren article on Bibliography page

48

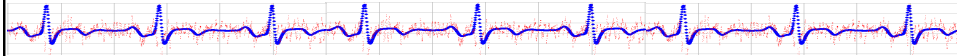
Refining a behavioral protocol



49

Behavior Coding Exercise

Groups of 2-4, one should have a laptop with R

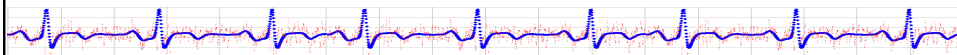


- You are developing a robotic couples counselor.
- You want to determine how couples react to it.
- Given nominal variable to code
- Discuss
 - Meaning
 - Refine values
 - Behavioral correlates
 - Draft coding manual
 - Focus on nonverbal behavior (poor audio)

51

Behavior Coding Exercise

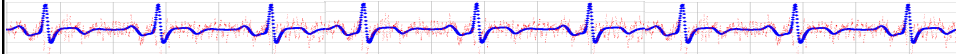
Groups of 2-4, one should have a laptop with R



- Shown 2-3 minute samples from 3 couples
- Interval sampling, 10s intervals
- Each judge codes behavior
 - Suggest shorthand, eg "E" for "Engaged"
 - Annotate Couple ID, landmarks (e.g. start of speaking turn), sample ID

52

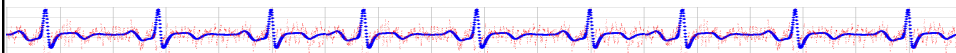
Code!



53

Behavior Coding Exercise

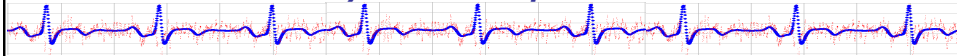
Groups of 2-4, one should have a laptop with R



- Put into one spreadsheet with one row per observation, one column per judge
- Compute interrater reliability
 - For >2 judges, compute mean of all pair-wise kappas
 - If <0.7 discuss discrepancies and improvements
 - Update coding manual
- Videos shown 2nd time for discussion
- Videos shown 3rd time for 2nd pass coding
- Repeat kappa calcs

54

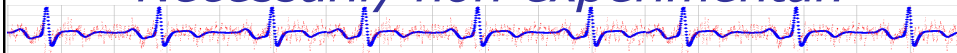
Other approaches to Data Collection *Necessarily non-experimental?*



- *Naturalistic Observation*
 - Unobtrusive observations of subjects' naturally occurring behavior are made
- *Ethnography*
 - The researcher becomes immersed in the behavioral or social system being studied. May be conducted as a participant or non-participant observation study
- *Sociometry*
 - You identify and measure interpersonal relationships within a group

58

Approaches to Data Collection *Necessarily non-experimental?*



- *Case History*
 - You observe and report on a single case
- *Content Analysis*
 - You analyze spoken or written records for the occurrence of specific categories of events (e.g., a word or phrase)

59

Approaches to Data Collection *Necessarily non-experimental?*



- *Archival Research*
 - You use existing records (e.g., police records) as your source of data
- *Meta-Analysis*
 - Compute overall statistics based on a number of previously-published studies.

60

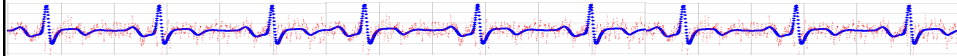
Sequential Analysis aka Time-Series Analysis



- B&A say recording sequences of behavior may yield more information than individual events.
 - e.g. interruption followed by grimace followed by rolling eyes

62

Content Analysis: Defining Characteristics

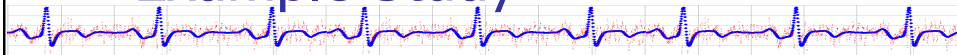


- Used to analyze a written or spoken record for occurrence of specific behaviors or events
- Archival sources often used as sources for data
- Response categories must be clearly defined
- A method for quantifying behavior must be defined

- Tools exist,
 - <http://www.lexicoder.com>
 - R package: <http://docs.quanteda.io/>

63

Example Study



- The CEO of Global Enterprises, Inc. is very worried about the low morale in the company, as evidenced by the amount of flame email she receives. She considers sending every office on a “ropes” course, but to do this would cost the company \$10M. She asks you to do a study to tell how well her scheme might actually work in reducing her flame mail.

64

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

65

Meta-Analysis

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

66

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?

67

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

68

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

69




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.

70




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.




71

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



72

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 $p = .14$.

73


APPENDIX A – EFFECT SIZES AND SIGNIFICANCE VALUES OF STUDIES INCLUDED

	Performance		Subjective		N
	Face vs. No Face	High vs. Low Realism	Face vs. No Face	High vs. Low Realism	
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,					50


■ Notes:

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

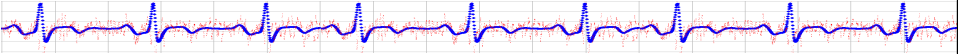
74



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 ($n = 18, r = .07$), $z = 2.43, p = .02$.



75

- ## Homework
- 
- Read Survey measures
 - B&A Ch 9
 - Article on debate re: scale questionnaire measures
 - Finish I3 (usability test)
 - Due next class

76