BRSM
Reliability & Outliers

Vinoo Alluri & Bapi Raju

Reliability

Reliability
/re-ly-a-bi-li-ti/
1. To be able to produce good res
time after time. 2. How much a
person can be depended on.

- **consistency** and **stability** of a research instrument (ex: measure or score or person)

- any measure we use in research should be reliable, otherwise it’s useless

- **repeatability** of a method/test or research findings

Kinds of Reliability

- Tools/methods or measuring device

- People

Threats to Reliability

- measurement error: equipment malfunction, human error, or ambiguous wording in survey questions
- instrumentation changes: measurement instruments are not consistent across repeated measurements, changes in the instrument itself can introduce variability and affect reliability.
- practice effects: Participants might improve their performance in a task due to practice or learning effects, leading to different results on subsequent administrations
- sampling variability: In experiments involving small sample sizes, random fluctuations in the characteristics of the participants can lead to unreliable results.

Threats to Reliability

- participant error: any factor which adversely alters the way in which the participant responds
  - ex: interview at 11 am vs 6 pm

- participant bias: any factor which produces a false/biased response
  - ex: mental health questionnaire in a company

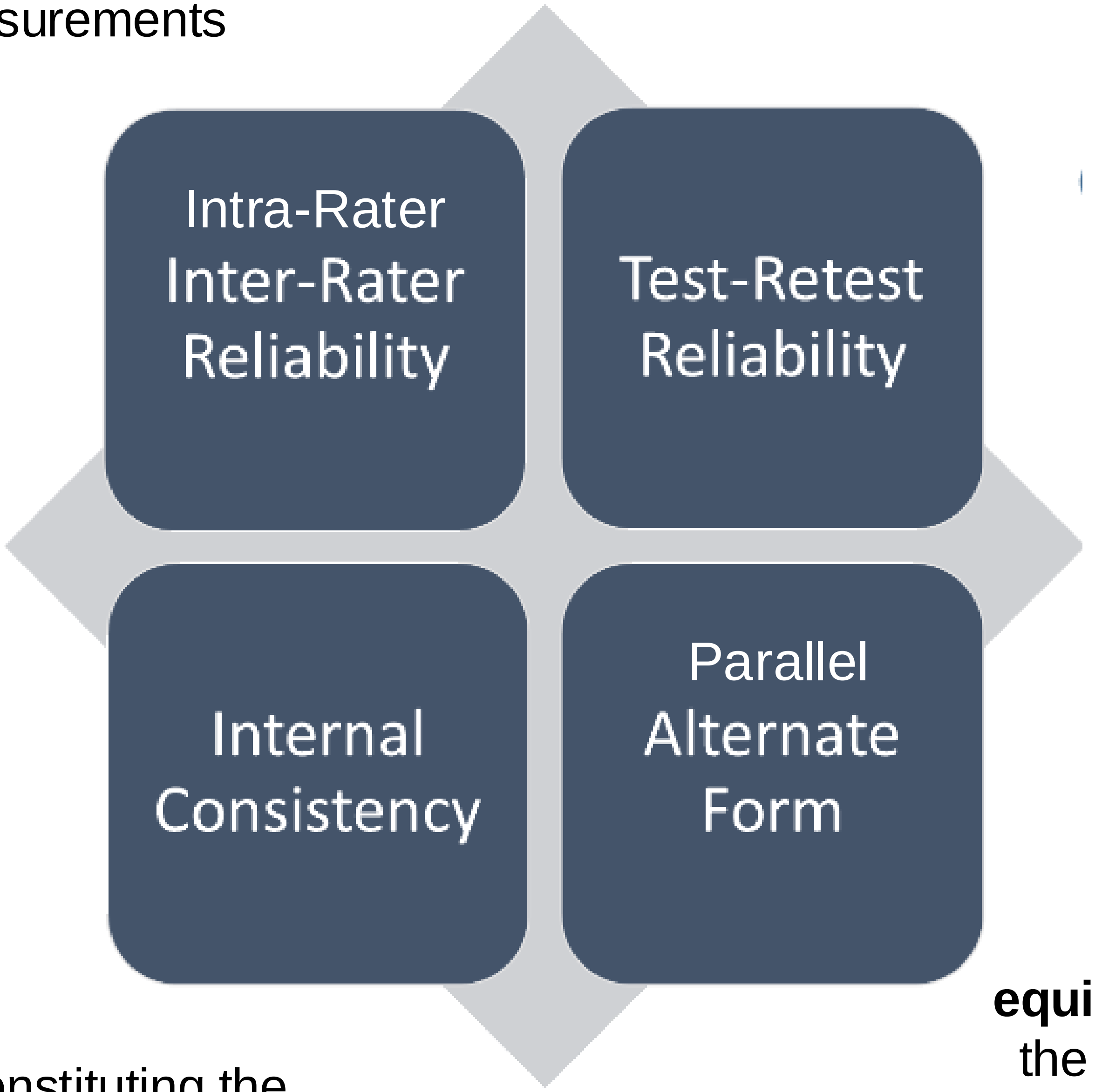
- researcher error: any factor which alters the researcher's interpretation
  - ex: fatigue effects if interview all day

- researcher bias: any factor which induces bias in the researcher's recording of responses
  - ex: subjective interpretation (to get the "result" you expect)

Kinds of Reliability

stability and **degree of agreement** between **people** during measurements

stability and **consistency** of method/tool/apparatus over time/repeated measurements



Intra-Rater
Inter-Rater
Reliability

Test-Retest
Reliability

Internal
Consistency

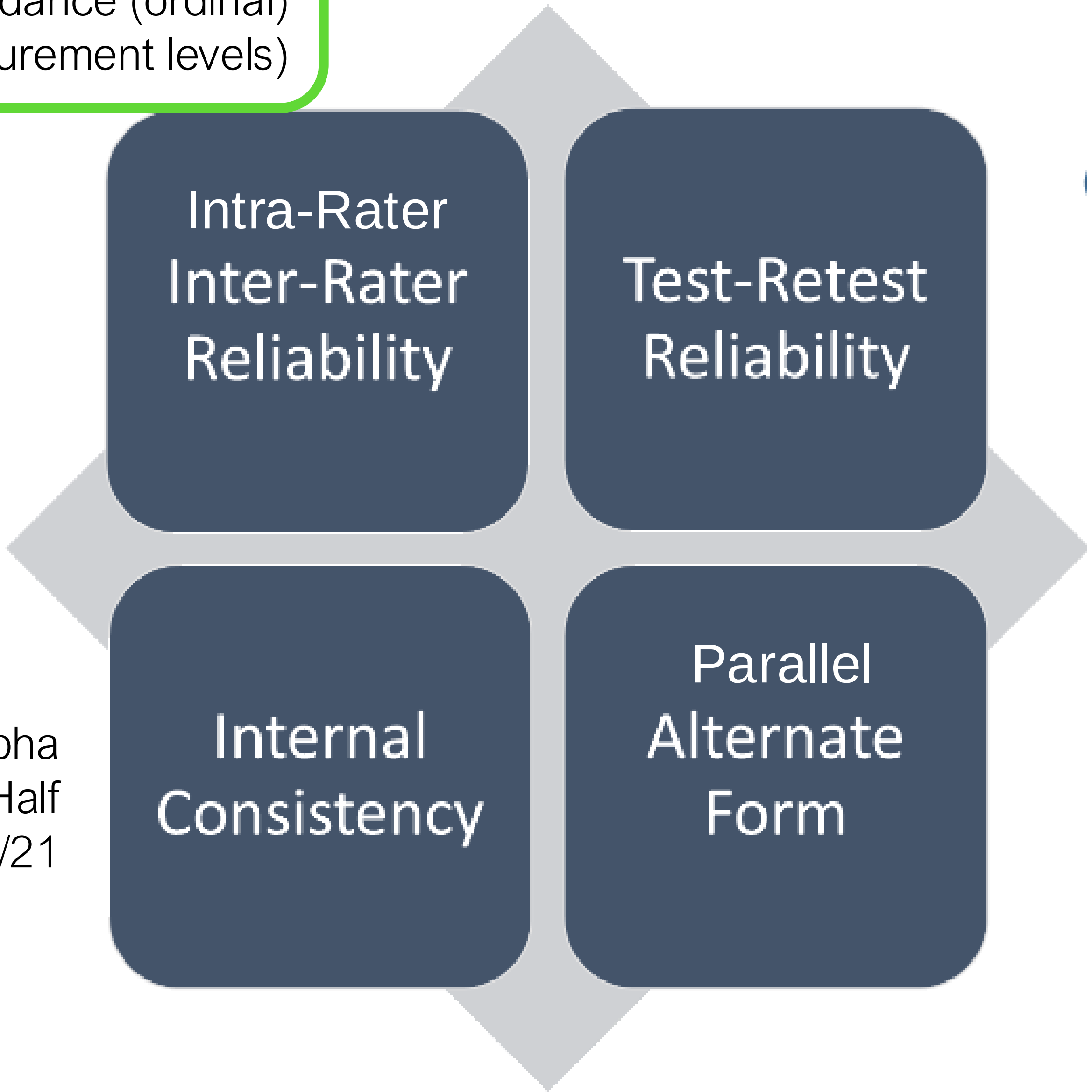
Parallel
Alternate
Form

coherence of attributes constituting the method/tool/apparatus

equivalence of two versions of the method/tool/apparatus to compare results

Kinds of Reliability

Cohen's Kappa (nominal; 2 raters)
Fleiss' Kappa(nominal; >2 raters)
Kendall's coefficient of concordance (ordinal)
Krippendorff's Alpha (all measurement levels)



Intra-Rater
Inter-Rater
Reliability

Test-Retest
Reliability

Internal
Consistency

Parallel
Alternate
Form

Pearson's correlation

Cronbach Alpha
Split-Half
Kuder Richardson-20/21

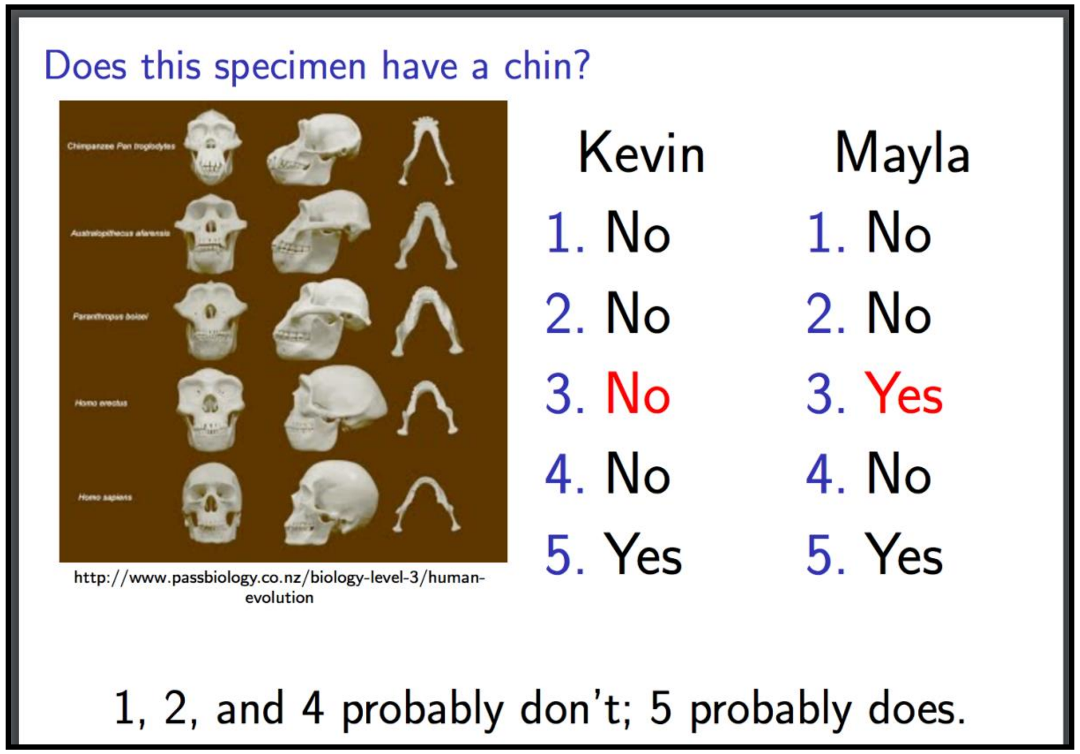
Reliability

- For people (reliability of participants)
  - Inter-rater or Inter-observer Reliability - degree of agreement between two participants or observers simultaneous recorded measurements
    - correlation, helps in outlier detection
  - Intra-observer Reliability - degree of agreement within the same observer’s measurements on repeated occasions

Reliability

EXAMPLE

- For people (reliability of participants)
  - Inter-rater or Inter-observer Reliability



Does this specimen have a chin?

Chimpanzee Pan troglodytes
Australopithecus afarensis
Paranthropus boisei
Homo erectus
Homo sapiens

http://www.passbiology.co.nz/biology-level-3/human-evolution

Kevin	Mayla
1. No	1. No
2. No	2. No
3. No	3. Yes
4. No	4. No
5. Yes	5. Yes

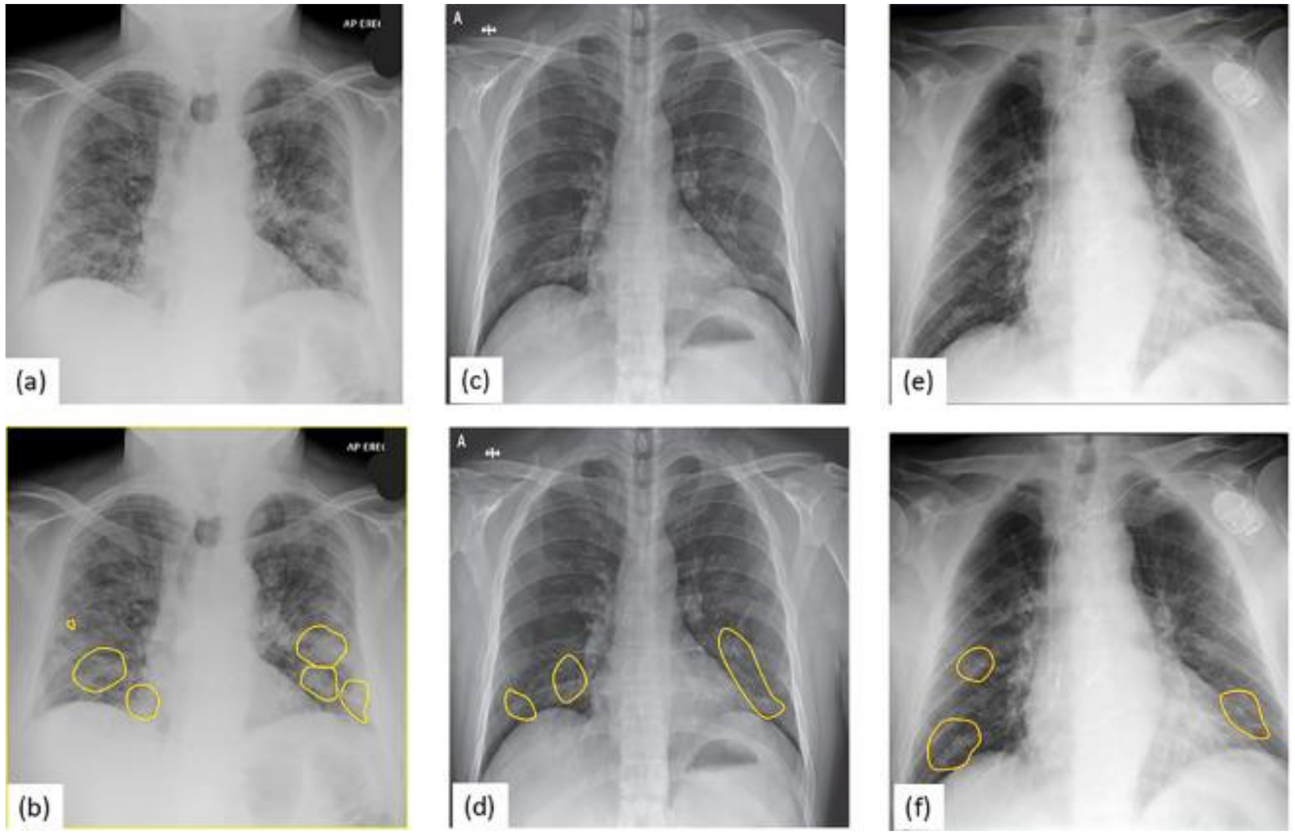
1, 2, and 4 probably don't; 5 probably does.

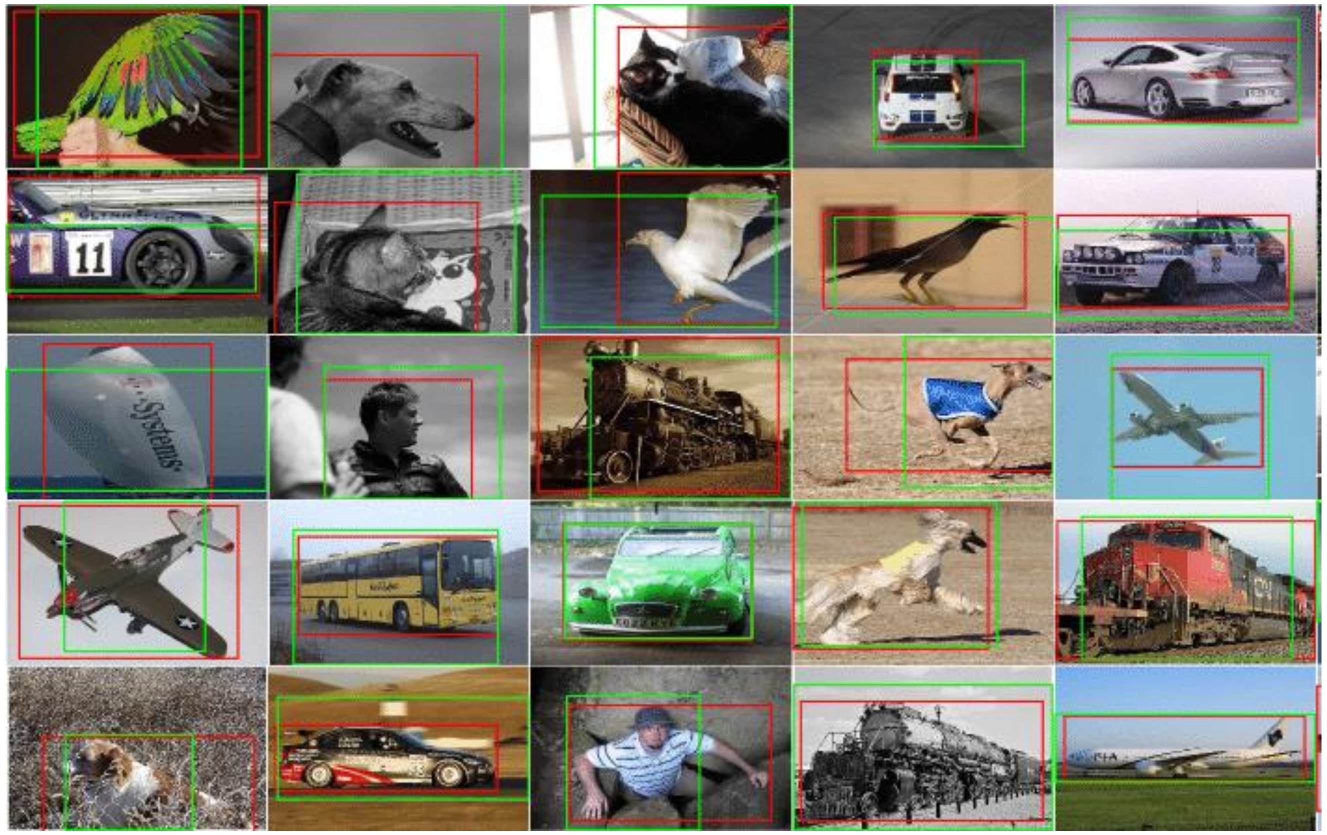
https://www.youtube.com/watch?v=fq_LNTPgVF8&app=desktop

Reliability

EXAMPLE

- For people (reliability of participants)
  - Inter-rater or Inter-observer Reliability

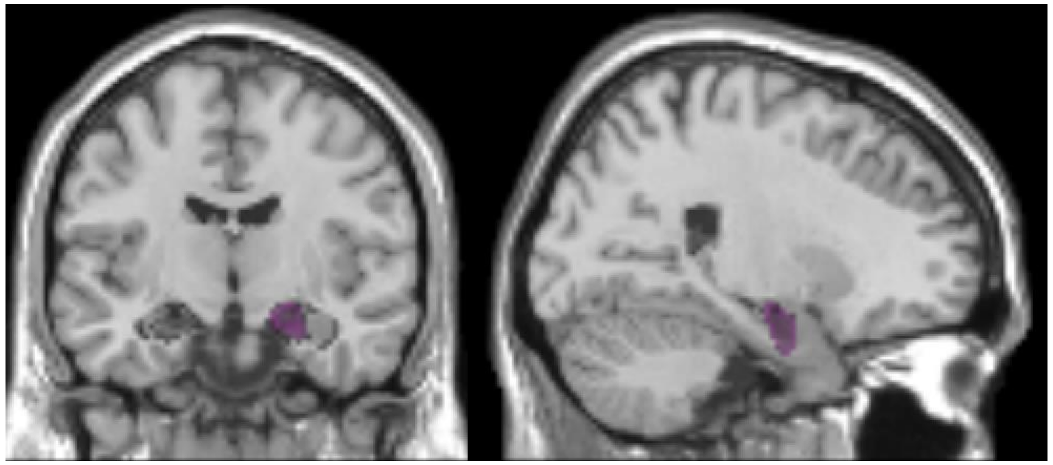




How many annotaters per dataset?

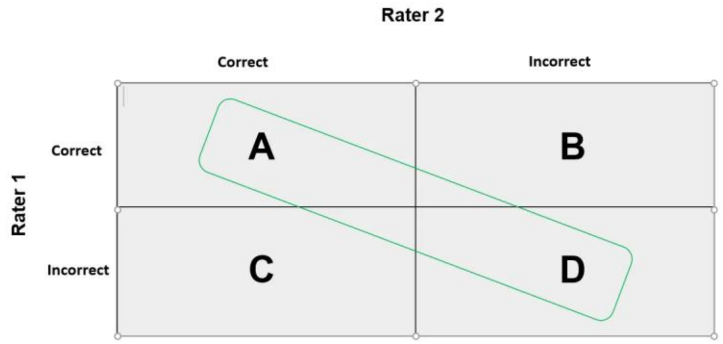
Reliability

- For people (reliability of researchers)
  - Similar to participants
  - not common
  - can be assessed in qualitative research when you have more than one PI
    - ex: qualitative thematic analysis



Reliability

- *Cohen's kappa*: a quantitative measure of reliability for two raters that are rating the same thing, correcting for how often the raters may agree by chance
- can be used to check consistency of the same rater at two different time points
- used when the variable is nominal



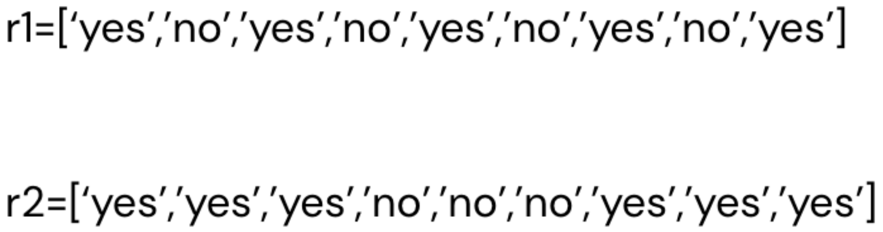
		Rater 2	
		Correct	Incorrect
Rater 1	Correct	A	B
	Incorrect	C	D

50 images rated by 2 raters

	Yes2	No2
Yes1	20	30
No1	35	15

Reliability

- *Cohen's kappa*: a quantitative measure of reliability for two raters that are rating the same thing, correcting for how often the raters may agree by chance



r1=['yes','no','yes','no','yes','no','yes','no','yes']

r2=['yes','yes','yes','no','no','no','yes','yes','yes']

Agreement= sum of agreements / total number of instances = (4+2)/9 = 0.66

	Yes2	No2
Yes1	4	1
No1	2	2

Reliability

– Internal consistency: Is the measurement device consistently measuring what you want it to measure?
  - Average inter-item correlation finds the average of all correlations between pairs of questions
  - Split Half Reliability: all items that measure the same thing are randomly split into two. The two halves of the test are given to a group of people and find the correlation between the two. The split-half reliability is the correlation between the two sets of scores.
  - Kuder-Richardson 20: average correlation for all the possible split half combinations in a test.

Reliability

- Internal consistency: Is the measurement device consistently measuring what you want it to measure?
  - *Cronbach's alpha*:
    - was developed in 1951 by Cronbach Lee to meet the need of finding an objective way of measuring the internal consistency reliability of an instrument used in a research work
    - mostly used when the research being carried out has multiple-item measures of a concept
    - typically used in questionnaires/surveys (self-reported)

Reliability

- Internal consistency: Is the measurement device consistently measuring what you want it to measure?
  - *Cronbach's alpha*:

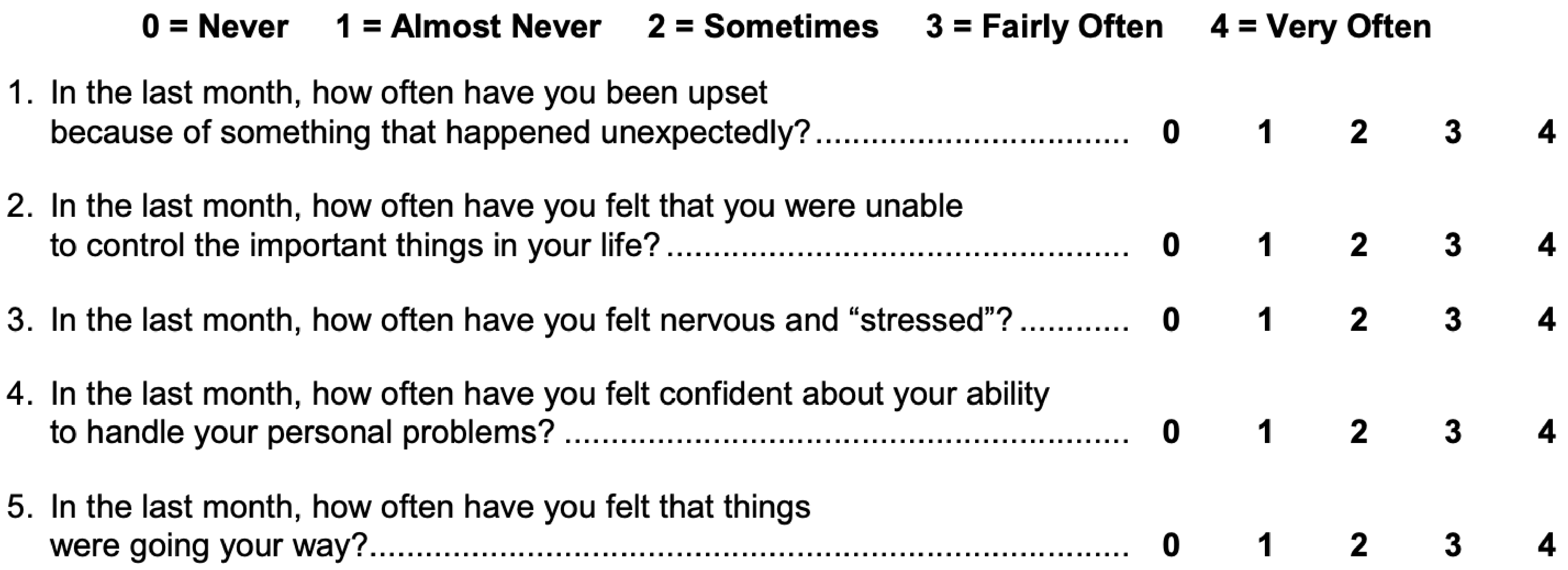
$$\alpha = \frac{k\bar{r}}{(1+(k-1)\bar{r})}$$

  - $\bar{r}$ = mean inter-indicator correlation
  - k=number of indicators or number of items

Reliability

EXAMPLE

- Internal consistency:
  - we have a 5 item scale showing data collected from 100 respondents



0 = Never   1 = Almost Never   2 = Sometimes   3 = Fairly Often   4 = Very Often

1. In the last month, how often have you been upset because of something that happened unexpectedly?................................. **0 1 2 3 4**
2. In the last month, how often have you felt that you were unable to control the important things in your life? **0 1 2 3 4**
3. In the last month, how often have you felt nervous and “stressed”? **0 1 2 3 4**
4. In the last month, how often have you felt confident about your ability to handle your personal problems? **0 1 2 3 4**
5. In the last month, how often have you felt that things were going your way?.. **0 1 2 3 4**

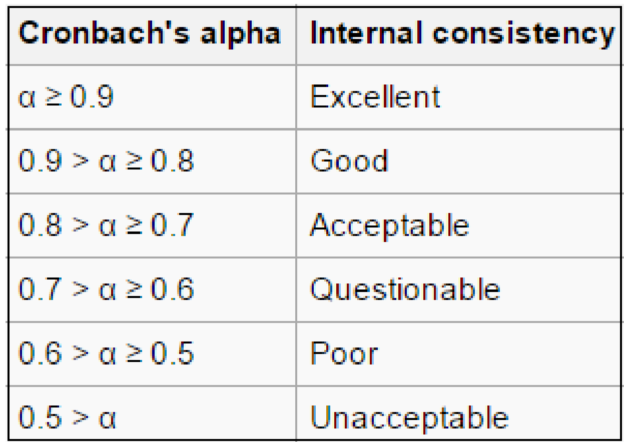
Reliability

EXAMPLE

- Internal consistency:
  - we have a 5 item scale showing data collected from 100 respondents
  - Correlate 100 responses x 5 items matrix

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1.0				
Item 2	.35	1.0			
Item 3	.42	.31	1.0		
Item 4	.25	.38	.41	1.0	
Item 5	.21	.36	.46	.31	1.0

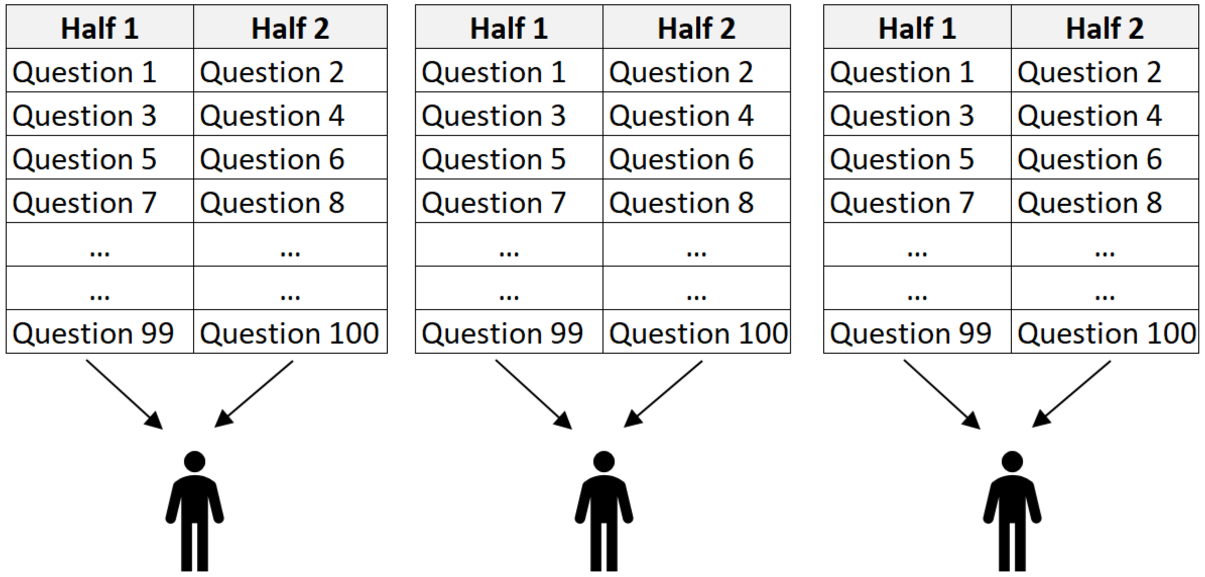
$$\alpha = \frac{k\bar{r}}{(1+(k-1)\bar{r})} = .73$$



Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent
0.9 > α ≥ 0.8	Good
0.8 > α ≥ 0.7	Acceptable
0.7 > α ≥ 0.6	Questionable
0.6 > α ≥ 0.5	Poor
0.5 > α	Unacceptable

Reliability

- Internal consistency: Is the measurement device consistently measuring what you want it to measure?
  - *Split-half* :
    - uses only some of available correlations;
    - compare results of one half to the other half.
    - If the test is reliable then people’s scores on each half should be similar

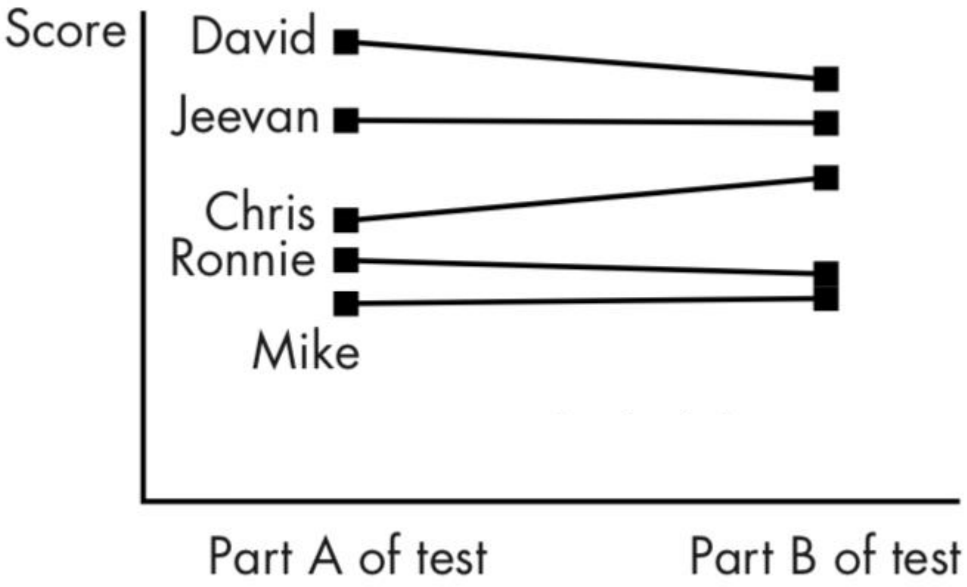


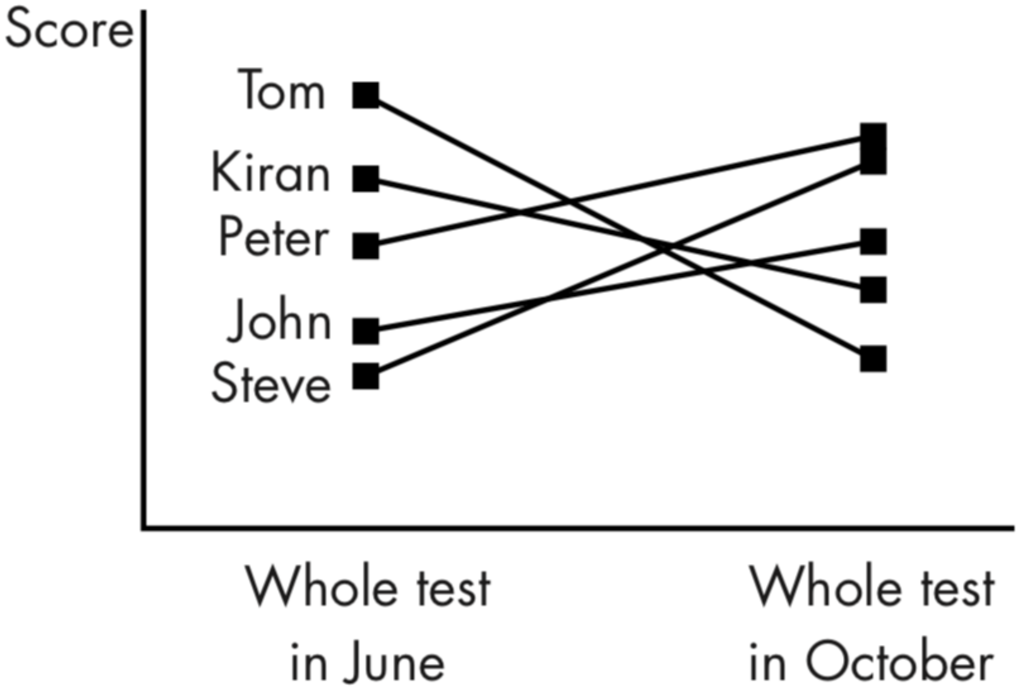
Half 1	Half 2
Question 1	Question 2
Question 3	Question 4
Question 5	Question 6
Question 7	Question 8
…	…
…	…
Question 99	Question 100

Half 1	Half 2
Question 1	Question 2
Question 3	Question 4
Question 5	Question 6
Question 7	Question 8
…	…
…	…
Question 99	Question 100

Half 1	Half 2
Question 1	Question 2
Question 3	Question 4
Question 5	Question 6
Question 7	Question 8
…	…
…	…
Question 99	Question 100

Reliability

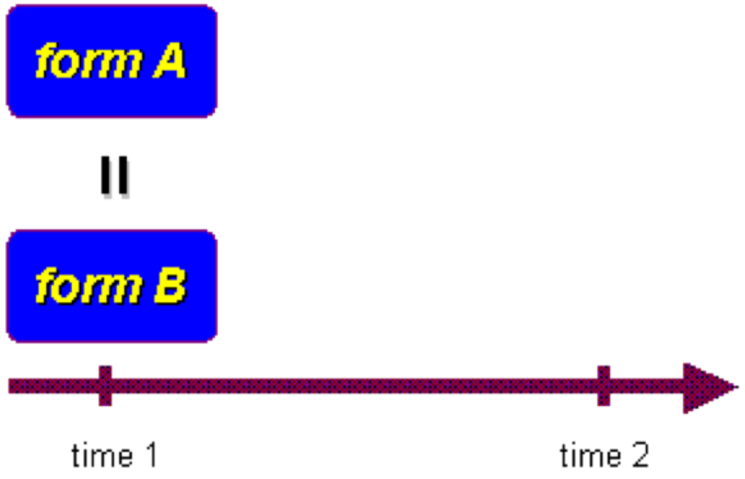


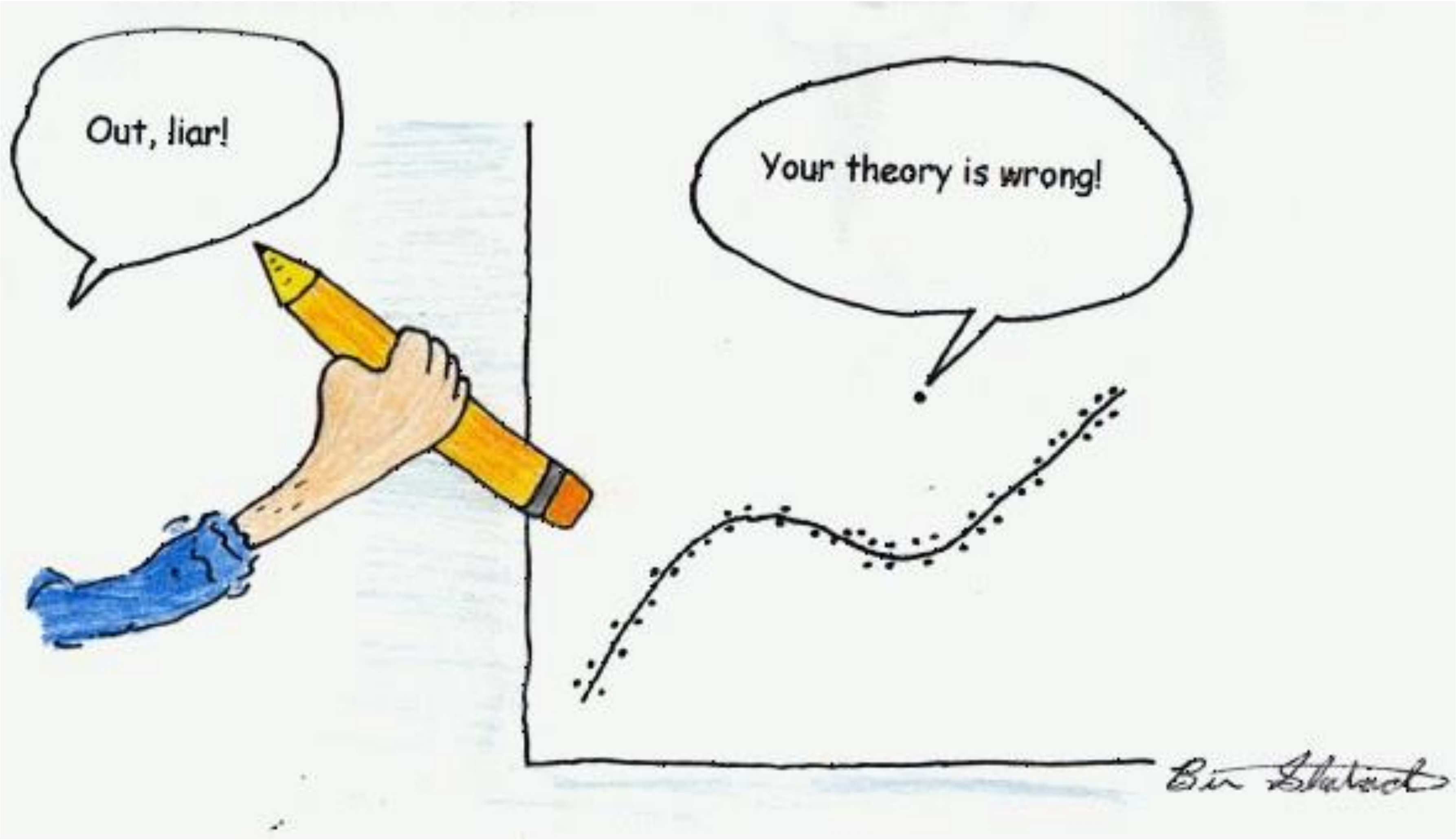


What kind of reliability and how good/bad is it?

Reliability

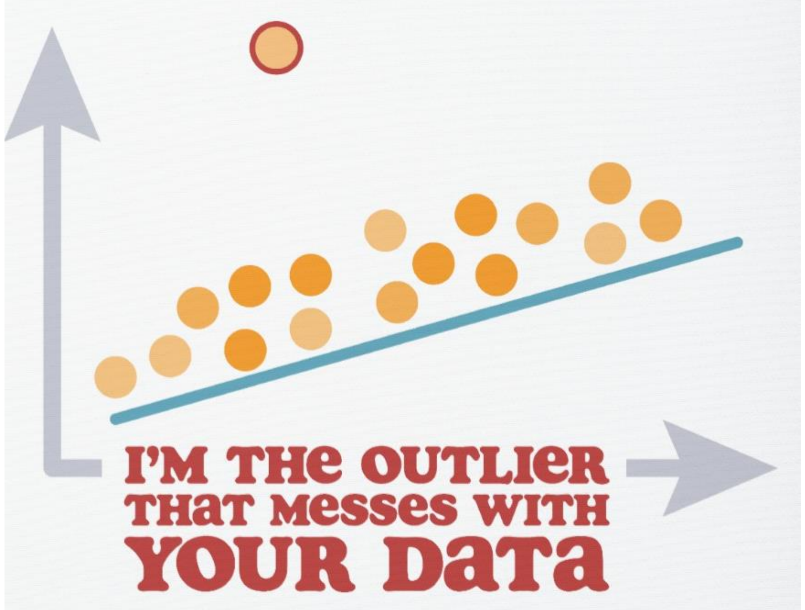
- parallel forms:
  - measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals
  - can avoid some problems inherent with test-resting





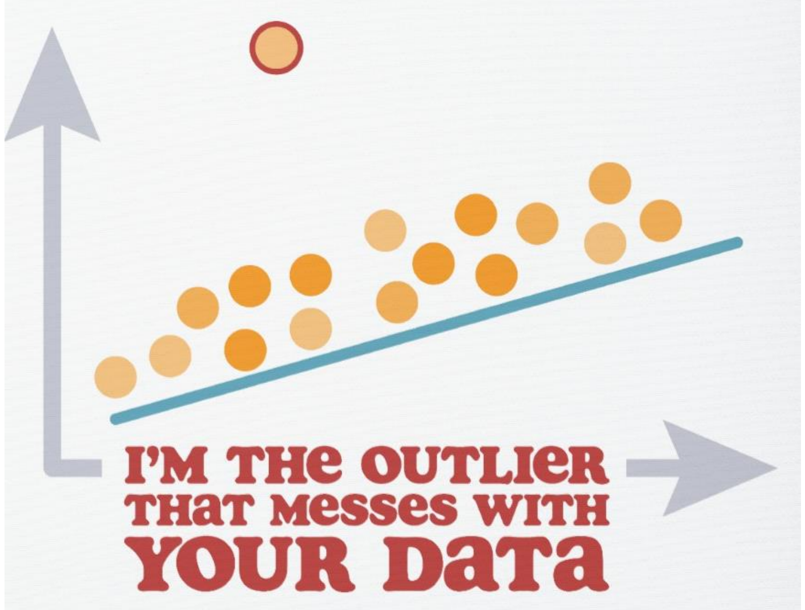
To have or not to have

Outliers



- detecting outliers is of major importance for almost any quantitative discipline (ie: Physics, Economy, Finance, Machine Learning, Cyber Security, Cognitive Science)
- not as common when sample size is low
  - ex: neuroimaging, qualitative studies involving interviews
- individual vs item/scale/stimulus

Outliers

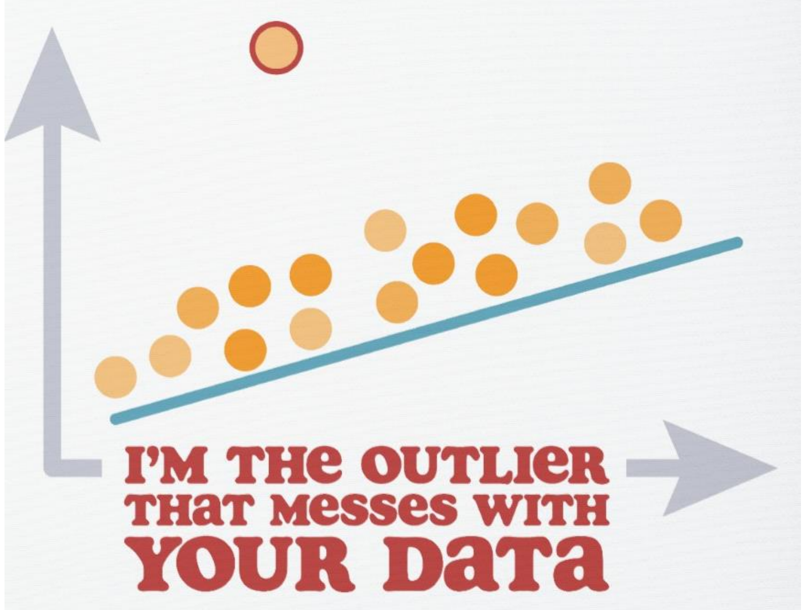


- probable causes?
  - measurement/execution errors (instrument errors/data extraction or experiment planning errors)
    - eg: improper scanner handling
  - data entry errors, missing data (human errors)
    - eg: entering 999 for missing values and using it for analysis

Dealing with Outliers

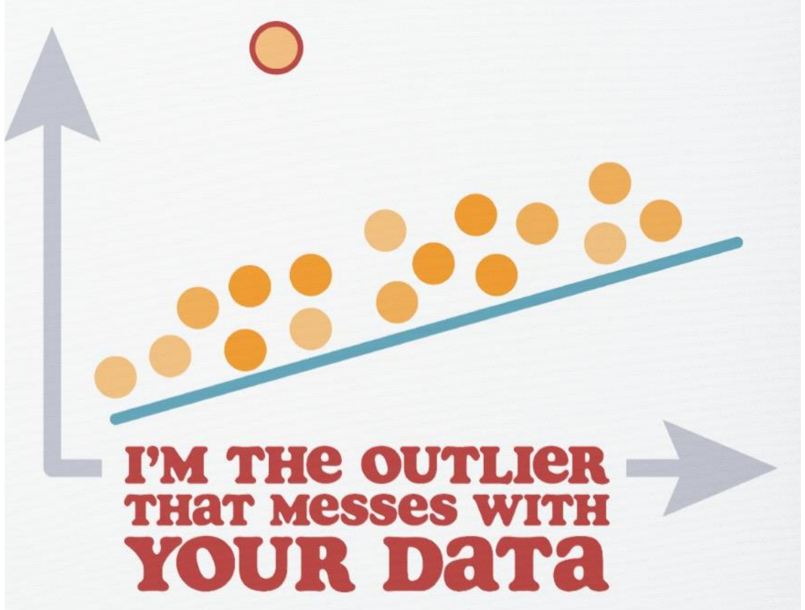
- omit
- replace (ex: with mean)
- using different analysis methods (ex: non-parametric tests)
- valuing the outliers
- data transformation

Outliers



- probable causes?
  - measurement/execution errors (instrument errors/data extraction or experiment planning errors)
    - eg: improper scanner handling
  - data entry errors, missing data (human errors)
    - eg: entering 999 for missing values and using it for analysis
  - data processing errors (data manipulation or data set unintended mutations)
    - eg: multiplying interval data

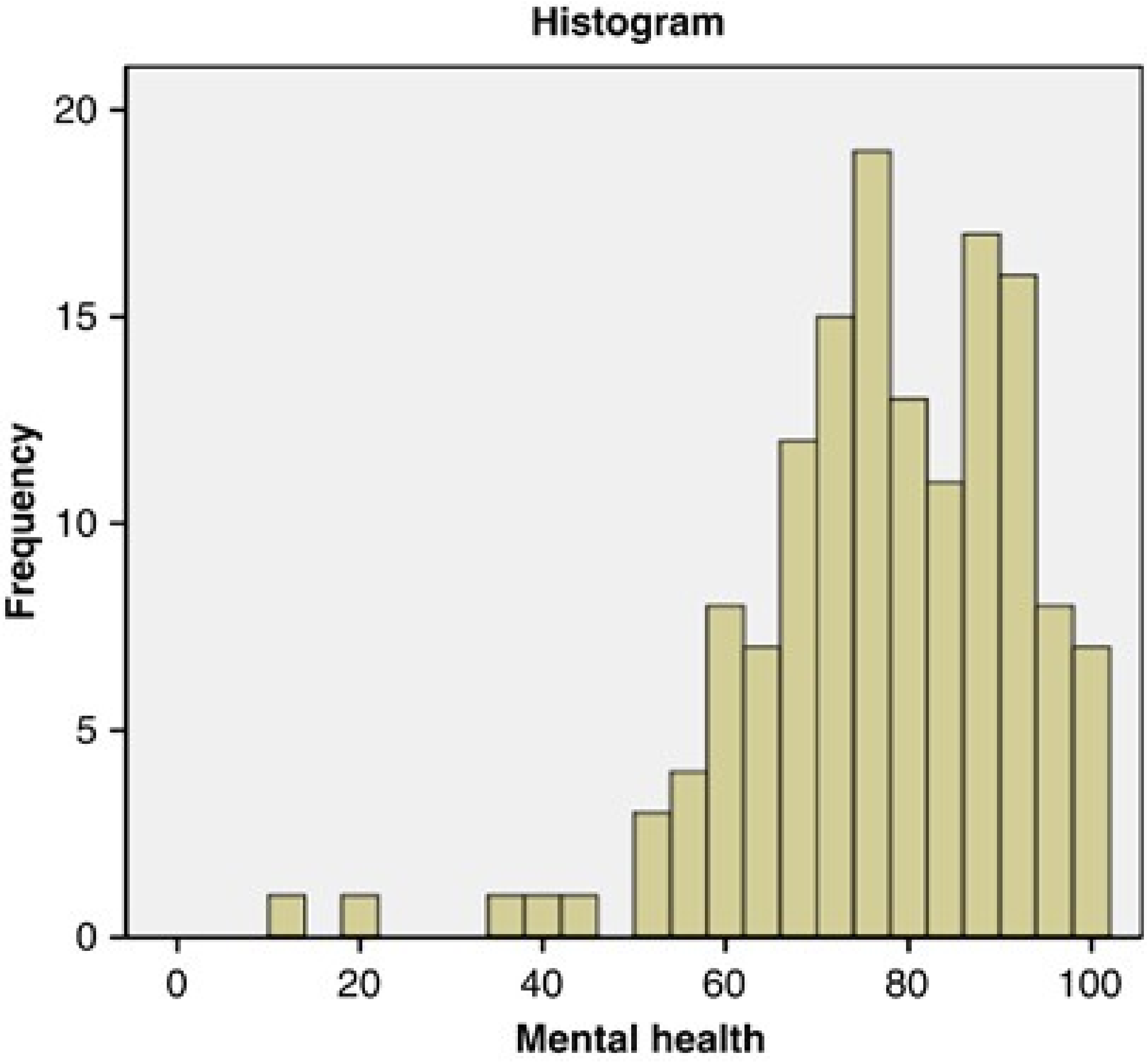
Outliers



- probable causes?
  - sampling errors (extracting or mixing data from wrong or various sources)
    - e.g: measure the weight of athletes but also include some wrestlers
  - natural (not an error, novelties in data or inherent data variability)

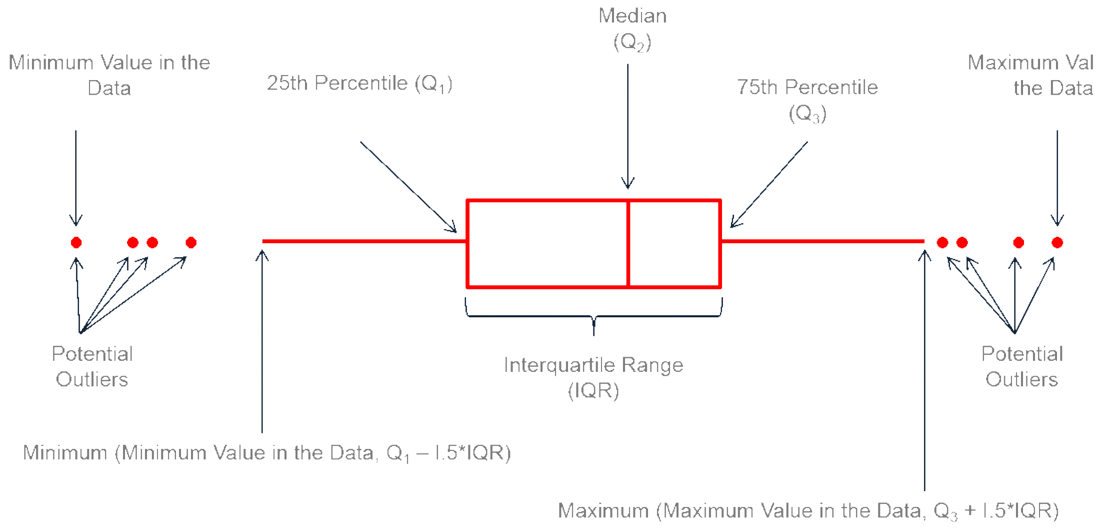
EXAMPLE

Natural Outliers



Outlier Detection

- graphical representations help (eg: scatter plot, box plot, histogram)



Median (Q_2)

Minimum Value in the Data

25th Percentile (Q_1)

75th Percentile (Q_3)

Maximum Value in the Data

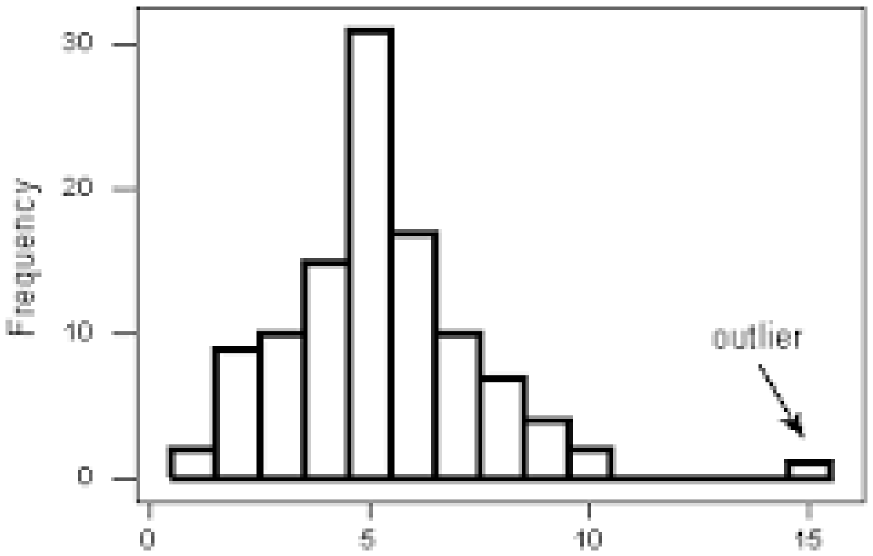
Potential Outliers

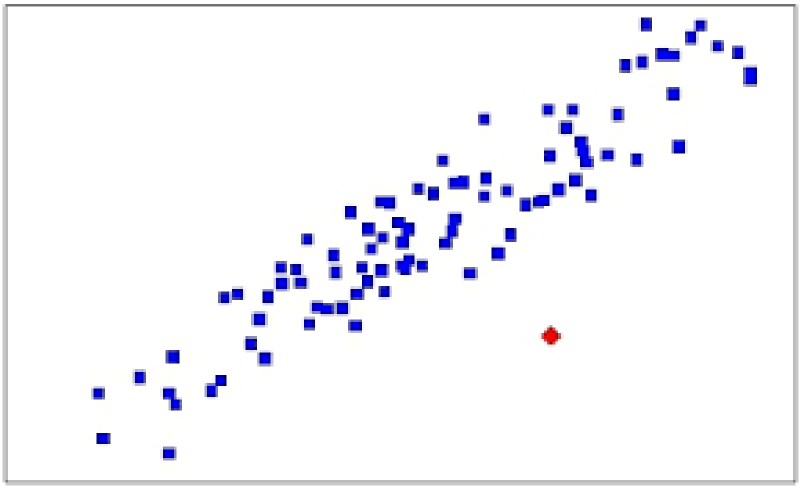
Interquartile Range (IQR)

Potential Outliers

Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)

Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)





Outlier Detection

Intuitive way of detecting outliers (esp. in a perceptual experiment or survey)?

Outlier Detection

- graphical representations help (scatter plot, box plot, histogram)
- >1.5 x InterQuartile Range
- 2/3 SDs from mean (depending on the nature of data)
- Grubbs’ test (single), Tietjen-Moore test (multiple), etc..

EXAMPLE

Outlier (individual) Detection

- 2/3 SDs from mean (depending on the nature of data)
  - check individual 2SDs away from mean rating of each

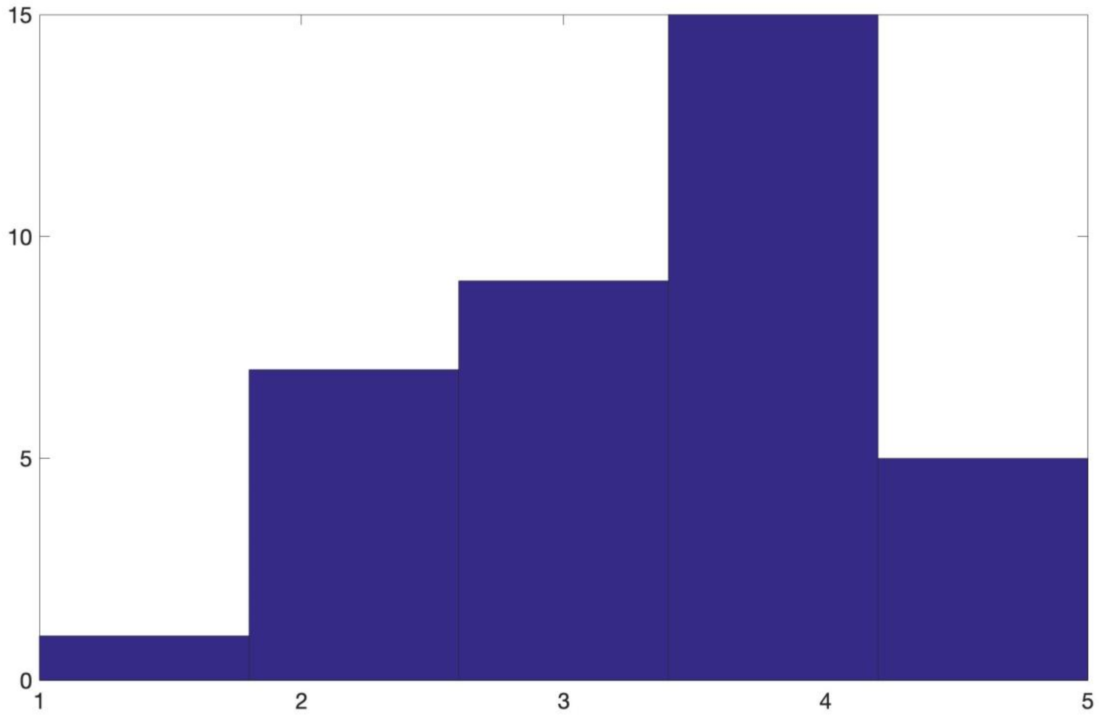
37 participants

37 x 100 Arousal ratings

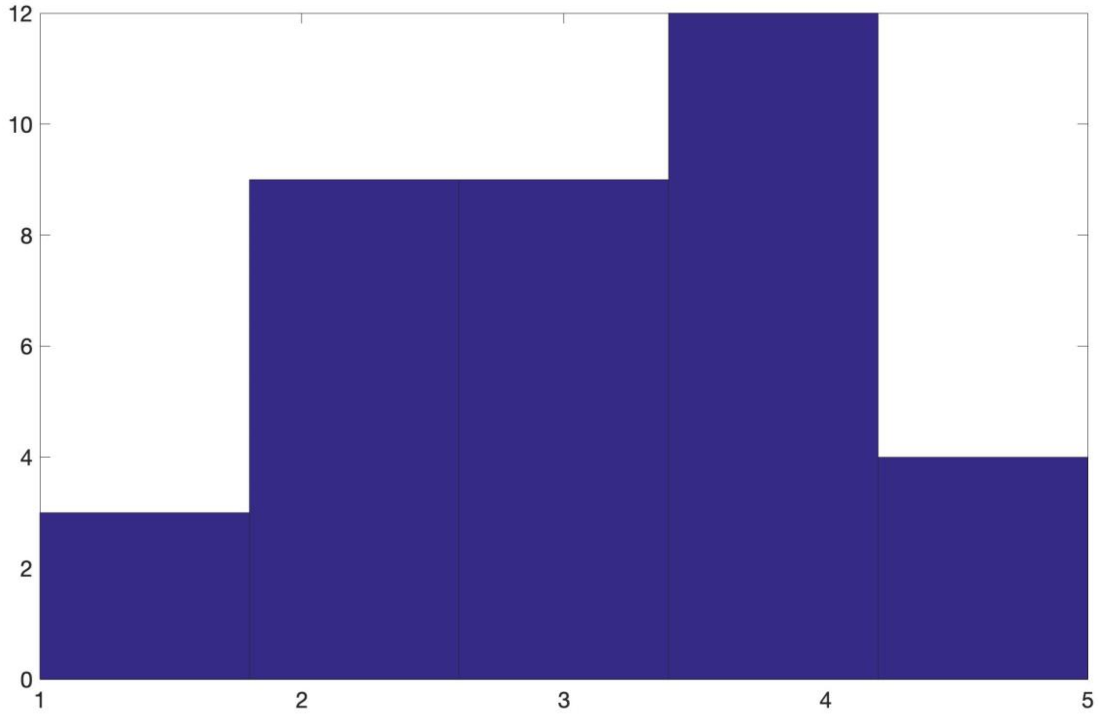
Rate Arousal (Energy) on a 5-point Likert scale
of
100 musical excerpts

EXAMPLE

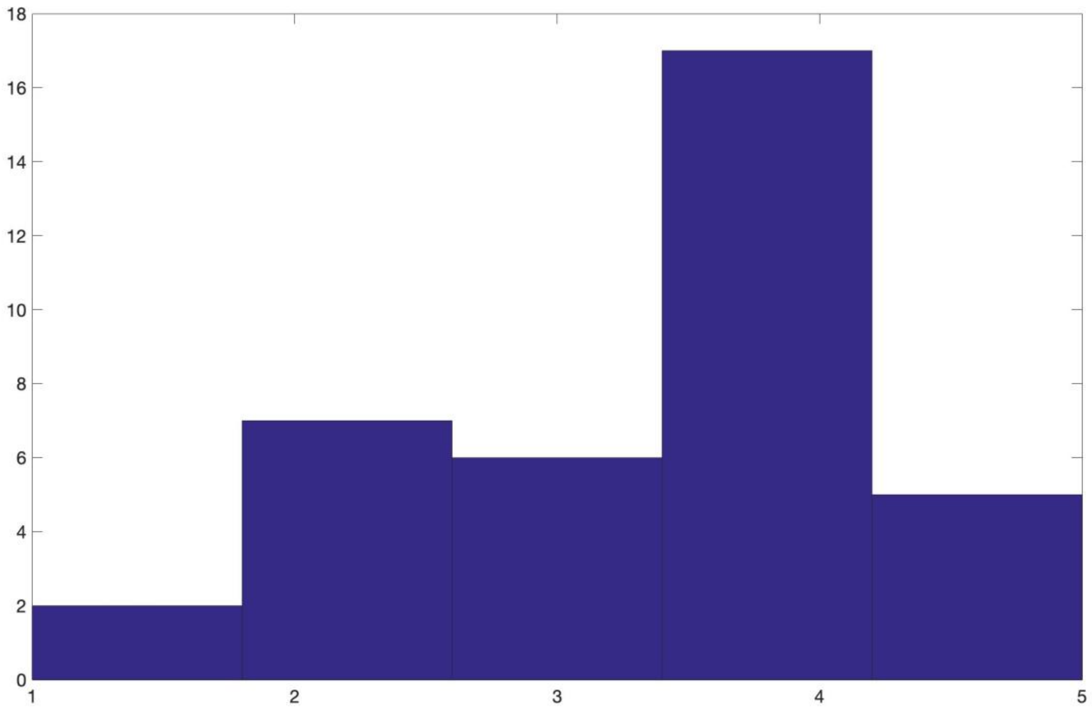
Outlier Detection



Stimulus 1 ratings



Stimulus 2 ratings



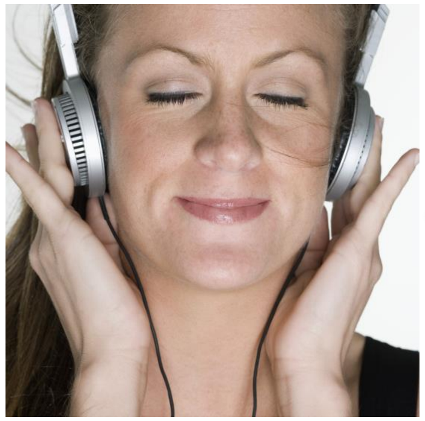
Stimulus 3 ratings

1 = low energy
5 = high energy

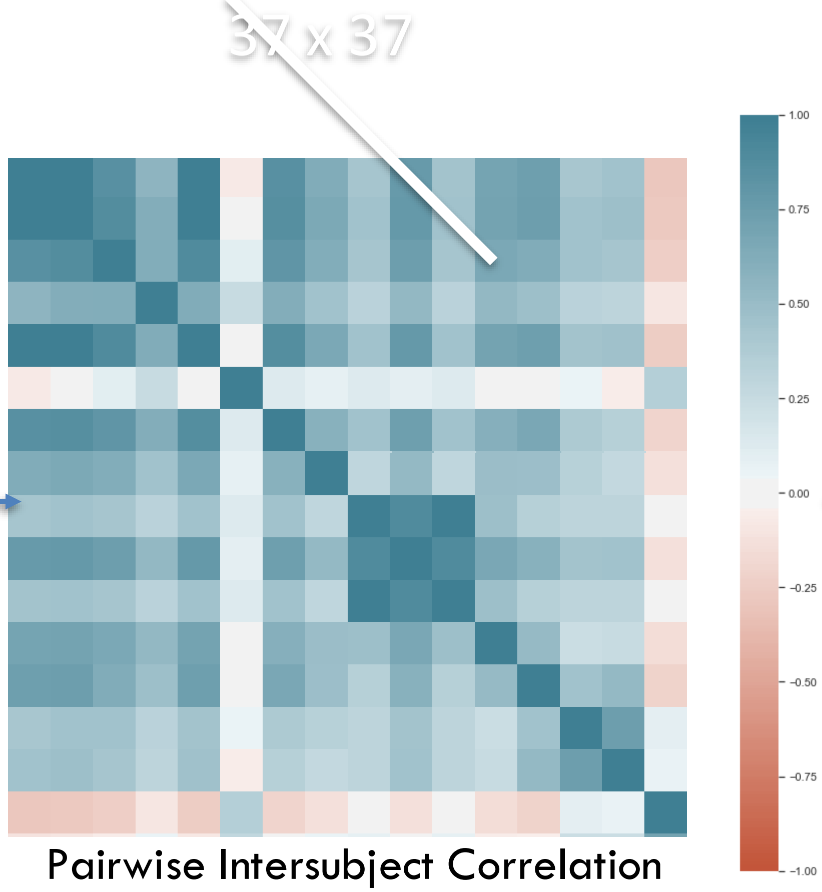
EXAMPLE

Outlier (individual) Detection

- 2SDs away from mean rating of each

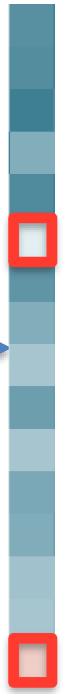


37 x 100 Arousal ratings



Pairwise Intersubject Correlation matrix

Mean ISC



EXAMPLE

Outlier (individual) Detection

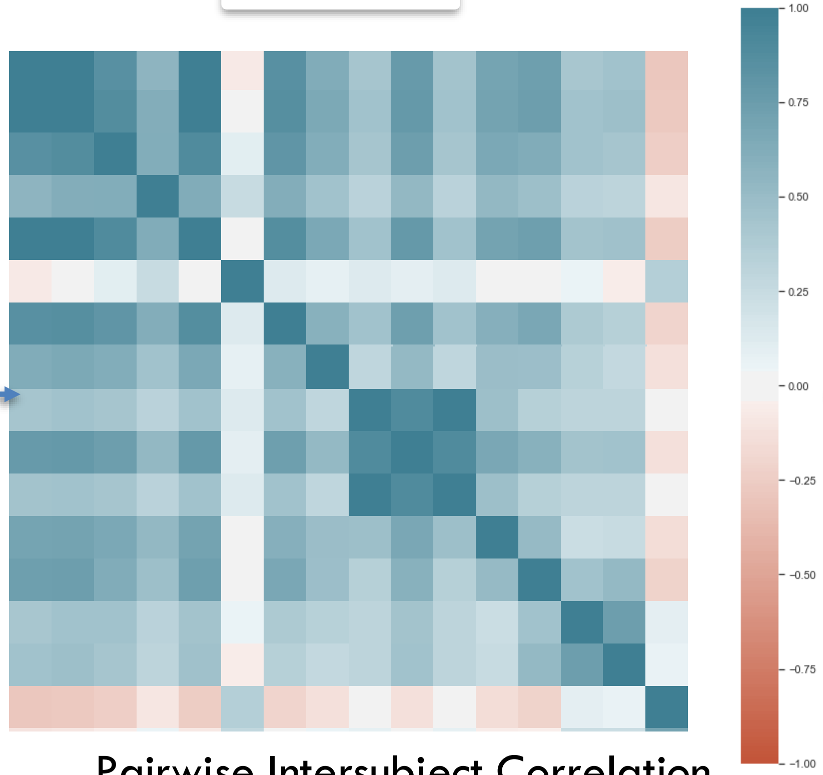
- 2SDs away from mean rating of each

29 x 29

29 participants

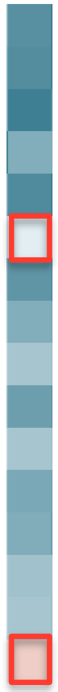
FOR SOMEONE YOUR BLOOD IS THE BEST HOLIDAY GIFT EVER
THANK YOU
MERRY CHRISTMAS & HAPPY NEW YEAR

Rate
Adventurous/Distressing of
100 advertisements



Pairwise Intersubject Correlation matrix

Mean ISC



not always suitable (especially for subjective ratings)!

Dealing with Outliers

- omit
- replace (ex: with mean)
- using different analysis methods (ex: non-parametric tests)
- valuing the outliers
- data transformation

Activity: Missing Values

- Omit
- Replace by frequent value (Mode)
- Replace by Mean / Median

Submit any 4 methods (names and 2-line description for estimating missing values!