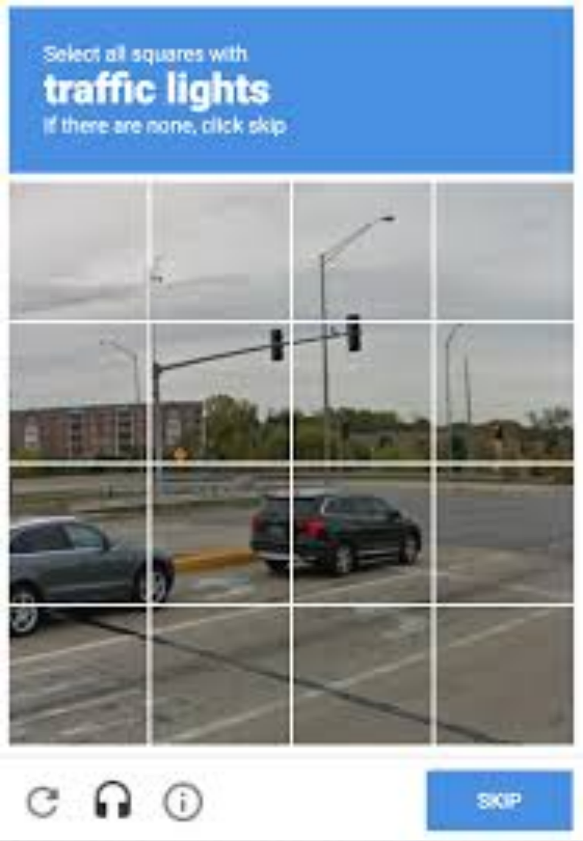
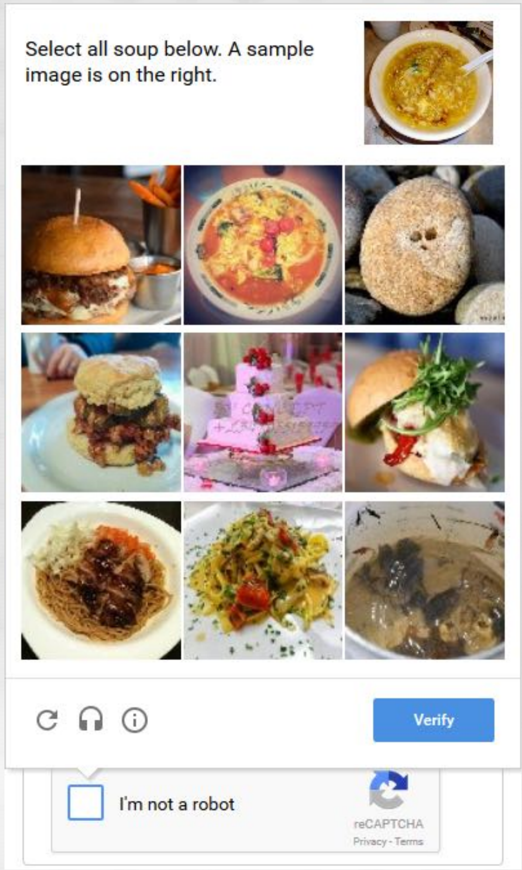

Select all squares with
traffic lights
If there are none, click skip


Select all soup below. A sample image is on the right.

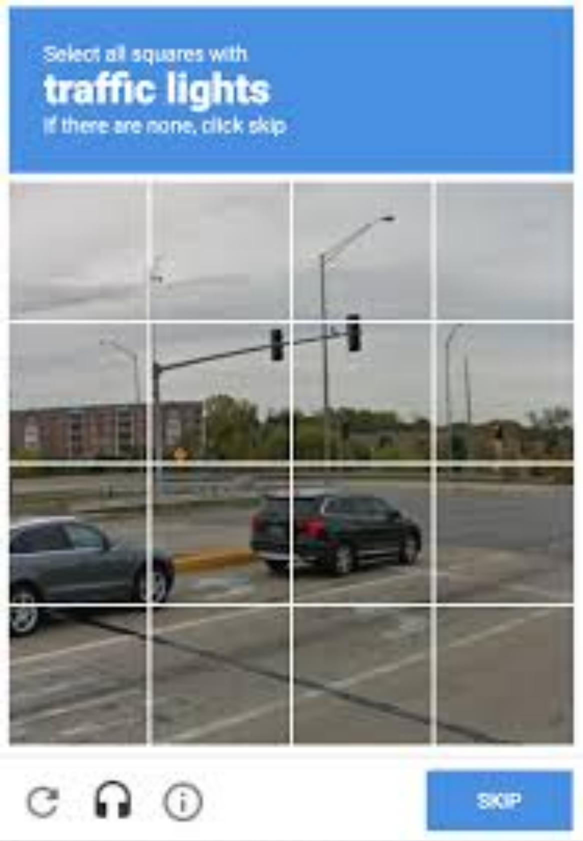
Coding and Reliability

DSA8022 Frontiers in Analytics

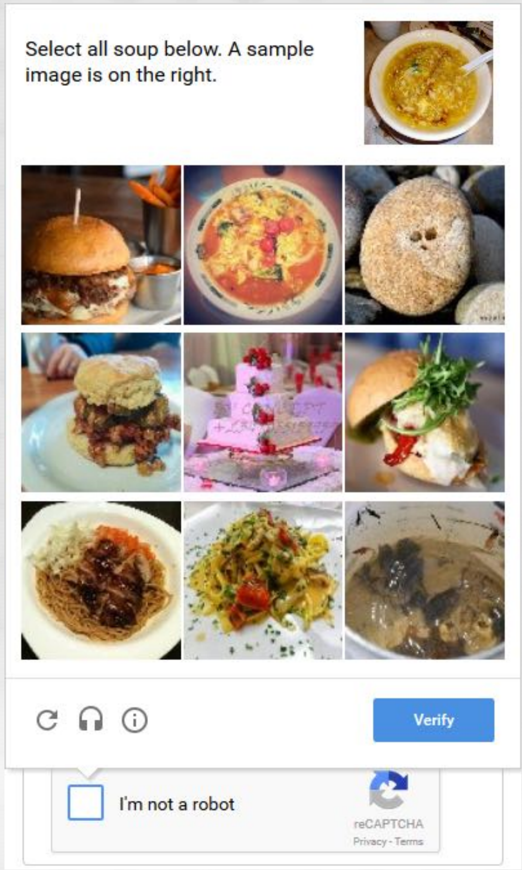
Dr Gary McKeown

School of Psychology, David Keir Building

0G3.536



Select all squares with
traffic lights
If there are none, click skip



Select all soup below. A sample image is on the right.

Coding and Reliability

DSA8022 Frontiers in Analytics

Dr Gary McKeown

School of Psychology, David Keir Building

0G3.536

OVERVIEW

- Supervised Learning
  - Creating a Ground Truth
  - Training Sets and Test Sets
- Observational Methods
- Inter-rater reliability
  - Categorical Data - Cohen’s Kappa
  - Continuous Data - icc
  - Both - Krippendorf’s alpha

SUPERVISED LEARNING

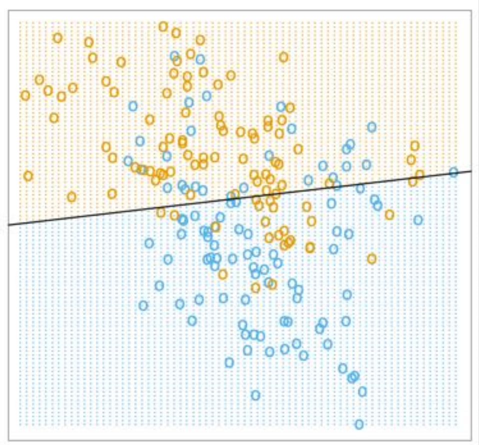
Ground Truth

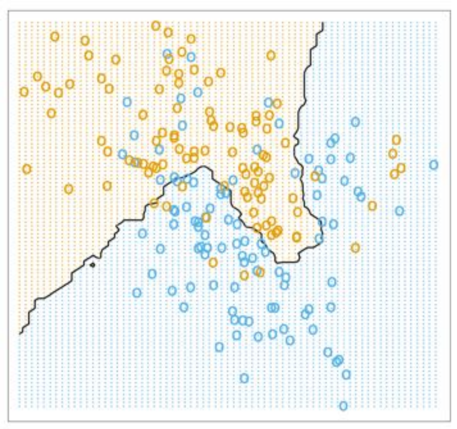
UNSUPERVISED LEARNING

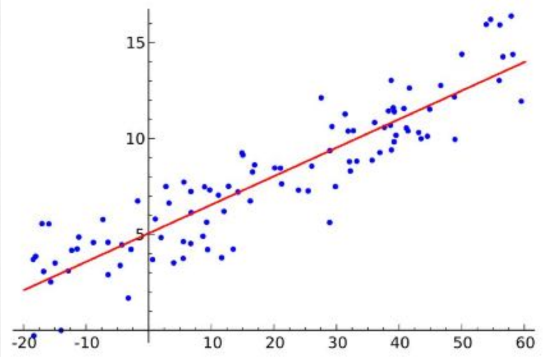
- Find the structure in the data without any explicit predefinition of the structure
  - Categorical Data
    - Clustering - k means
    - Topic modelling - Latent Dirichlet Allocation (LDA)
  - Continuous Data
    - Dimensionality Reduction
    - e.g Bag of Words Sentiment Analysis
    - Latent Semantic Analysis

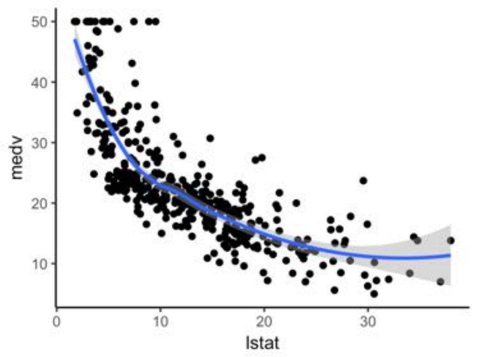
SUPERVISED LEARNING

- Give a ground truth that defines the data and learn the relationship
- Training and test datasets
  - Categorical Data
    - Classification
  - Continuous Data
    - Regression





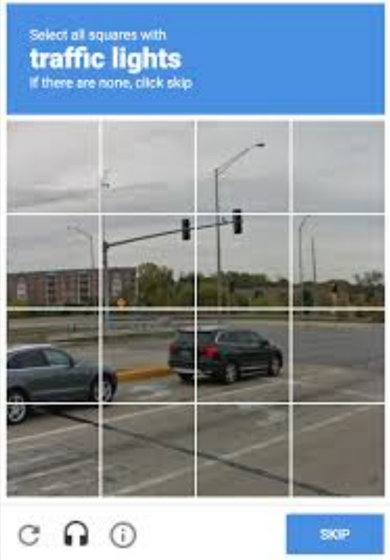




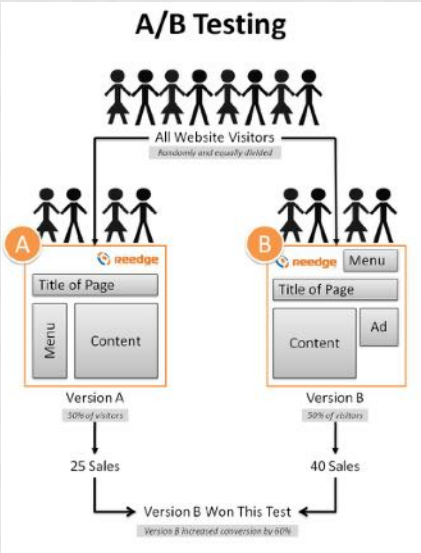
OBSERVATIONAL METHODS

MEASUREMENT

- We can measure things from may different sources
- Surveys
- Self-report
- A/B Testing
- Getting lots of people to rate things and tell us what they are
- Strategies
  - Lots of naive raters
  - Smaller numbers of expert raters



Select all squares with
traffic lights
If there are none, click skip



A/B Testing

All Website Visitors
Randomly and equally divided

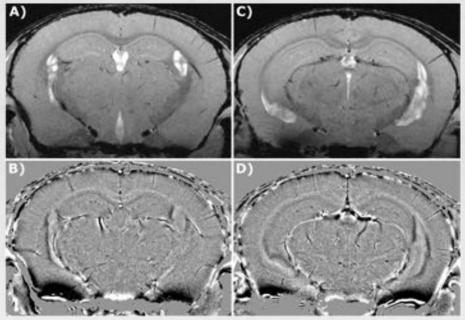
Version A
50% of visitors
25 Sales

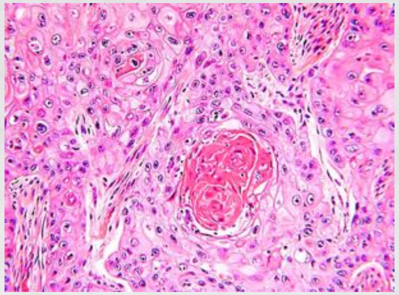
Version B
50% of visitors
40 Sales

Version B Won This Test
Version B increased conversion by 60%

Amazon Mechanical Turk

Access a global, on-demand, 24x7 workforce





MEASUREMENT

- Lab settings
  - Specific highly controlled tasks
  - Lower ecological validity
  - Overfitted
  - Does not generalise
- In the wild
  - Natural behaviour
  - Uncontrolled
  - High ecological validity
  - More noisy

INTERRATER RELIABILITY

RELIABILITY

Observed Score = True Score + Measurement Error

- Observed Score is typically the Score on a Psychometric test
- Reliability tries to estimate the proportion of variance that is captured by any measurement tool
- Typically scores are between 0 and 1.
- A reliability of 0.8 means 80% is True Score 20% is Error variance
- High reliability means we have a better measure of what it is we are trying to measure - it is closer to the true score
- Assuming error is independent - random not systematic

MEASURING RELIABILITY

Nominal Binary

Subject	Rater 1	Rater 2
1	Yes	No
2	Yes	Yes
3	No	No
4	No	Yes
5	Yes	Yes
6	No	No
7	No	Yes

Nominal Multiple Categories

Subject	Rater 1	Rater 2
1	Cat3	Cat3
2	Cat2	Cat4
3	Cat2	Cat2
4	Cat1	Cat1
5	Cat4	Cat2
6	Cat3	Cat2
7	Cat4	Cat4

MEASURING RELIABILITY

Ordinal

Subject	Rater 1	Rater 2
1	Cat3	Cat3
2	Cat2	Cat4
3	Cat2	Cat2
4	Cat1	Cat1
5	Cat4	Cat2
6	Cat3	Cat2
7	Cat4	Cat4

Continuous

Subject	Rater 1	Rater 2
1	66.43	65.33
2	45.98	54.88
3	67.98	66.98
4	76.99	64.87
5	34.55	56.72
6	78.99	57.91
7	76.88	89.43

MEASURING RELIABILITY

- How to measure the relationship between raters?
  - Pearson’s correlation?
  - Percentage agreement?

- Issues of chance agreement
- Issues of systematic error

Subject	Rater 1	Rater 2
1	66.43	65.33
2	45.98	54.88
3	67.98	66.98
4	76.99	64.87
5	34.55	56.72
6	78.99	57.91
7	76.88	89.43

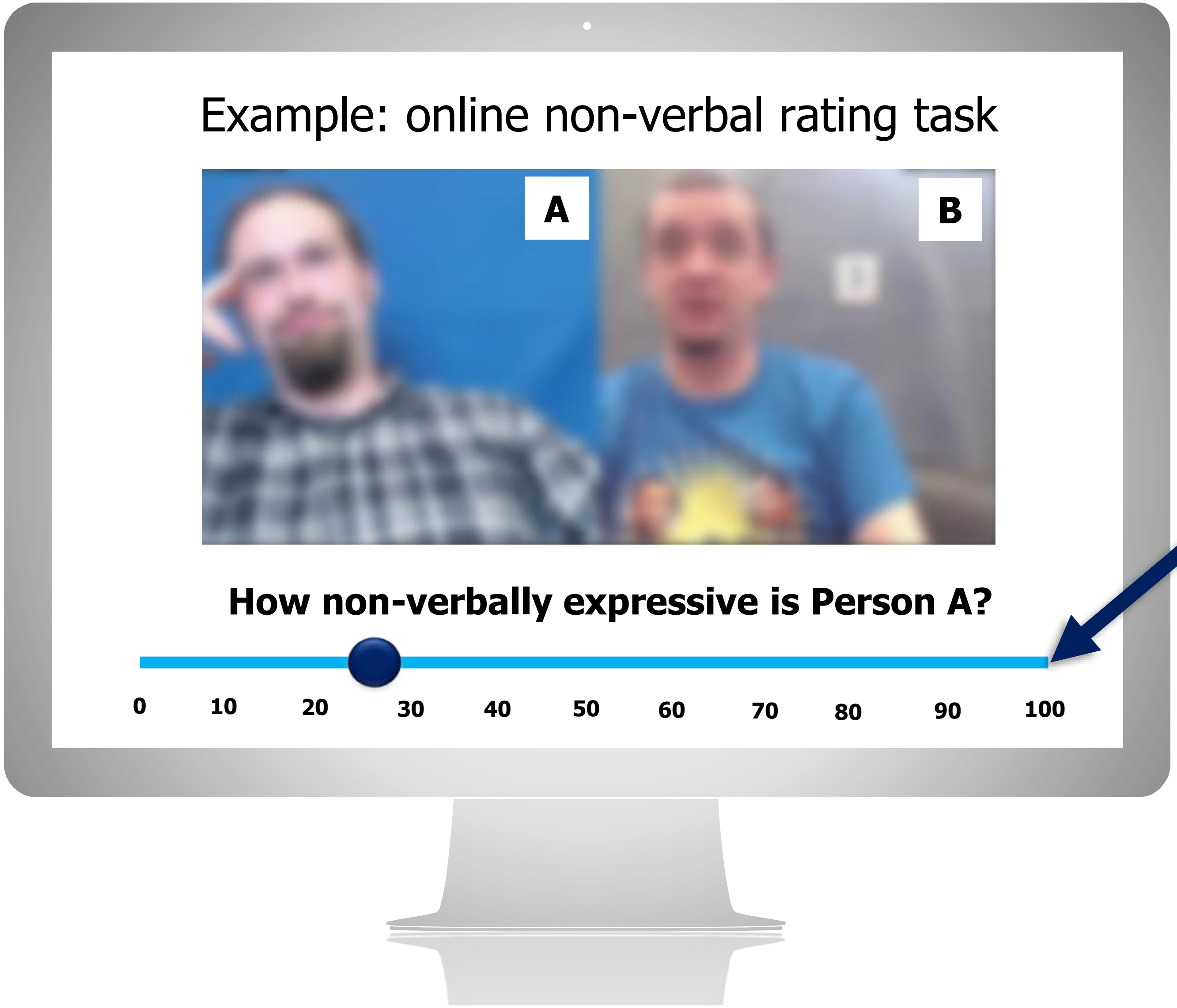
INTER RATER RELIABILITY

- Quantifies the degree of agreement between two or more raters who make independent ratings about the features of a set of subjects.
- Subjects can be people, behaviours, images, social media posts, adverts, movie reviews
- Many options that depend on:
  - The nature of the data
  - The number of raters
  - The goal of the rating
  - The amount of ratings that have been conducted - full coverage or partial coverage

INTER RATER RELIABILITY

- Many options:
  - Percentage agreement
  - Cohen's Kappa
    - Nominal data - 2 raters
  - Cohen's weighted Kappa
    - Ordinal data - 2 raters
  - Fleiss' Kappa
    - Nominal data - 2 or more raters
  - Intra Class Correlation
    - Continuous/interval data - multiple raters
  - Krippendorf's alpha
    - Nominal, Ordinal, Interval data, many raters

An example of behaviour analysis - non-verbal expressivity slider measurement on a continuous scale



Example: online non-verbal rating task

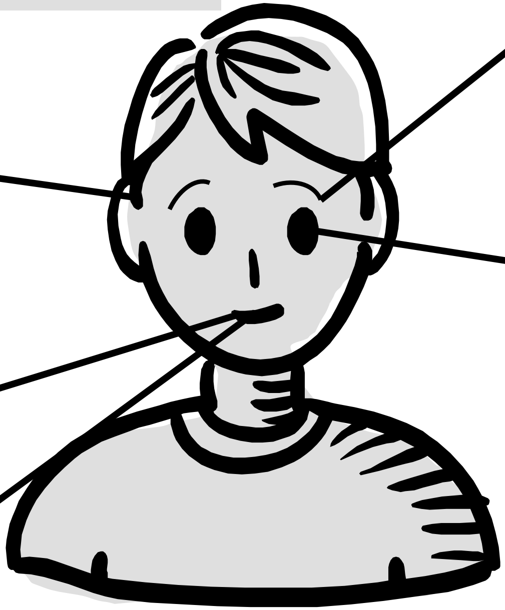
How non-verbally expressive is Person A?

Moveable, marked slider.

0 to 100.

2 decimal points.

Micro-level nonverbal annotation



head

- Nodding
- Cock to side
- Shaking
- Other positional

smiling

- Duration
- Intensity 1 - 3

scowling

- Duration
- Intensity 1 - 3

eyebrows

- Raise
- Furrow
- Other

eyes

- Gaze direction
- Mutual gaze
- Exaggerated opening
- Closing

hand

- Self-touch
- Self-adaptors
- Object-adaptors
- Gestures

Vocalisation/verbalisation behaviours

- All audible sounds
- Speaking duration
- Interruptions
- Back-channelling
- Laughter duration
- Laughter intensity

THIN SLICING IN TIME

Non-verbal behaviour score

Time (seconds)	Rater 1	Rater 2	Rater 3
1-15	66.43	65.33	64.11
16-30	45.98	54.88	67.35
31-45	67.98	66.98	67.35
46-60	76.99	64.87	76.23
61-75	34.55	56.72	61.87
76-90	78.99	57.91	65.98

COHEN'S KAPPA

- Percentage agreement
  - Does not take chance into account
- Cohen's Kappa
  - The original measure by Jacob Cohen
  - Accounts for chance

$$\kappa = 1 - \frac{1 - p_o}{1 - p_c}$$

p_o is the relative observed agreement

p_c probability of chance agreement

- Limitations
  - Can be biased in certain circumstances
  - Only nominal data - Only 2 raters
  - Requires fully-crossed designs

COHEN'S WEIGHTED KAPPA

- There are many variations on Cohen's Kappa
- Cohen's Weighted Kappa
  - Provided by Cohen
  - The weights are the penalties given rot different ways of disagreeing
  - The weighting makes it suitable for ordinal data.
  - Many possible weightings
- Limitations
  - Only 2 raters
  - Equivalent to two-way mixed, single-measures, consistency ICC (just use ICC?)

FLIESS' KAPPA

- There are many variations on Cohen's Kappa
- Fleiss' Kappa
  - Provided by Joseph Fleiss
  - Two or more raters
  - Raters are ample from a larger population and new ones are sampled for each subject
- Limitations
  - Only nominal data
  - Assumes a new sample of raters for each subject
  - Therefore not useful for fully crossed designs
    - Might be good for recaptcha traffic lights

INTRACLASS CORRELATION

- ICC - Intraclass correlation
- Flexible
  - Can be used for ordinal, interval or ratio data
  - Two or more coders
  - Can do fully crossed where all subjects are rater by multiple raters
  - Also not fully -crossed where multiple rater rate a subset and a single rater rates the rest
  - Values between 0 and 1
    - 1 equals higher reliability/agreement
- Limitations
  - Many forms can be confusing (6 or more forms)
  - Some implementations deal poorly with missing data

INTRACLASS CORRELATION

- Hallgren (2012) provides a nice tutorial
- Four decisions need to be made before running an ICC:

- Which model: One-way or Two-way?
- Absolute agreement or consistency?
- Unit of Analysis: Average-measures or Single-measures?
- Random or fixed effects?

Open ICCData.csv

Using the irr package
conduct an intra-class correlation

You will need to know the options:

Oneway or twoway?

Absolute agreement or consistency?

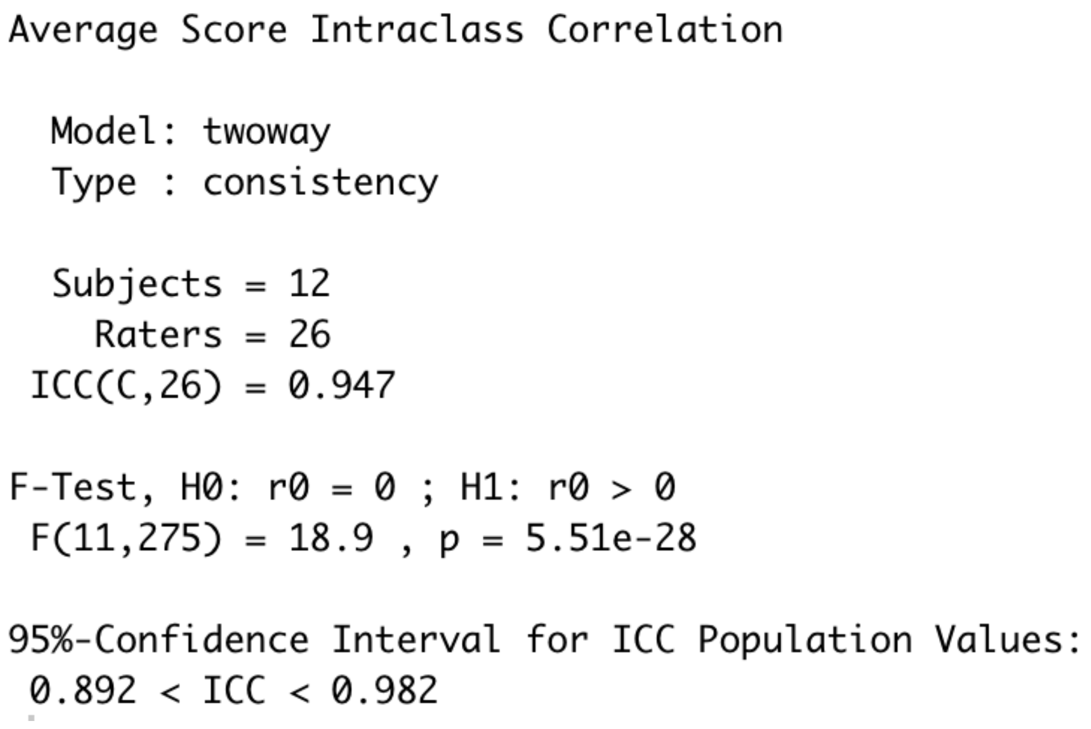
Average measures or single measures?

Random or fixed effects?

You can answer two of these from looking directly at the data can you work out which of the two need more information?

Two-way, consistency, average

```r
icc_result <- icc(ICCData, model="twoway", type="consistency",unit="average")
```



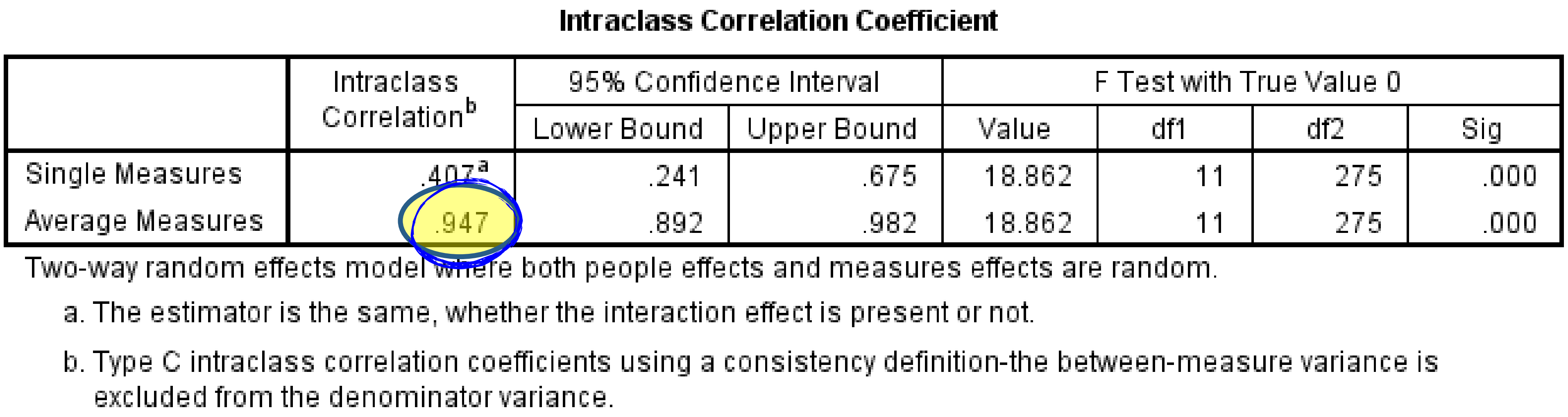
```
Average Score Intraclass Correlation

  Model: twoway
  Type : consistency

  Subjects = 12
    Raters = 26
 ICC(C,26) = 0.947

F-Test, H0: r0 = 0 ; H1: r0 > 0
 F(11,275) = 18.9 , p = 5.51e-28

95%-Confidence Interval for ICC Population Values:
 0.892 < ICC < 0.982
```



Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.407[a]	.241	.675	18.862	11	275	.000
Average Measures	.947	.892	.982	18.862	11	275	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

McGraw and Wong (1996)

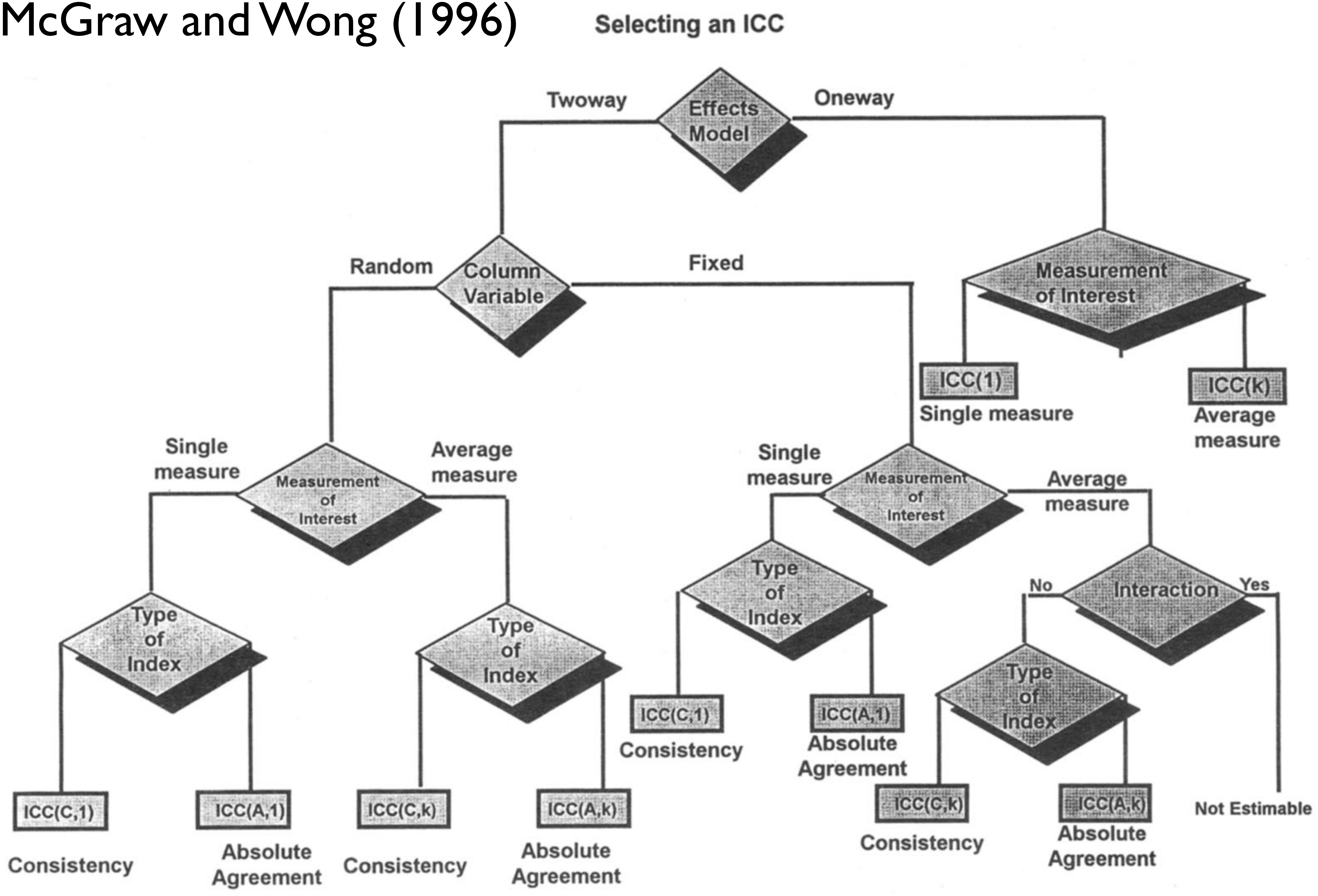
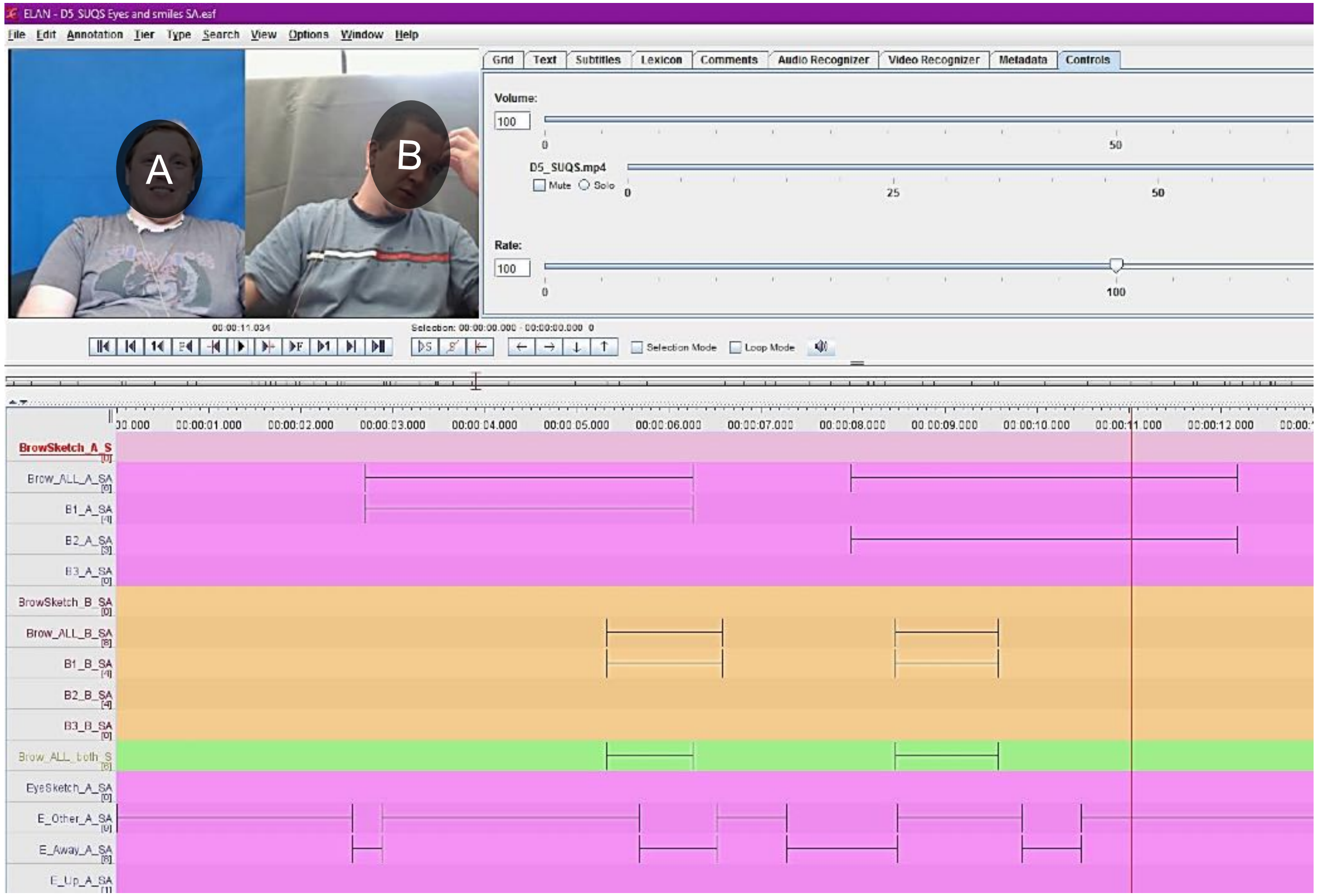


Figure 1. Flow chart for selecting an appropriate intraclass correlation coefficient (ICC).

KRIPPENDORF'S ALPHA

- The most flexible in terms of data types
- Works for nominal, ordinal, interval and ratio data
- Good at handling missing data
- Values between 0 and 1
  - 1 equals higher reliability/agreement
- Allows comparison across data types
- Limitations
  - Has not been widely adopted
  - Less flexible than ICC for continuous data

ELAN manual annotation software


ELAN - D5_SUQS Eyes and smiles SA.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comments Audio Recognizer Video Recognizer Metadata Controls
Volume:
100
D5_SUQS.mp4
□ Mute ○ Solo
Rate:
100
A
B
00:00:11.034
Selection: 00:00:00.000 - 00:00:00.000 0
□ Selection Mode □ Loop Mode
BrowSketch_A_S
Brow_ALL_A_SA
B1_A_SA
B2_A_SA
B3_A_SA
BrowSketch_B_SA
Brow_ALL_B_SA
B1_B_SA
B2_B_SA
B3_B_SA
Brow_ALL_both_S
EyeSketch_A_SA
E_Other_A_SA
E_Away_A_SA
E_Up_A_SA

MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

The Language Archive

http://tla.mpi.nl/tools/tla-tools/elan/

Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands