Inter-Rater Reliability

Jeffrey M. Girard
Carnegie Mellon University

Inter-Rater Reliability

- What role does reliability play in affective computing?
  - Measurements are often used to train supervised learning algorithms
  - These training labels are called “ground truth” and *assumed* to be correct
  - The reliability of training labels can impact algorithm performance
  - Any biases inherent to the labels will likely be inherited by the algorithm
  - If you want to interpret a variable, you should estimate its reliability

Inter-Rater Reliability

- How can reliability be estimated or explored?
  - Similar (or identical) objects are measured in different contexts
    - Have multiple observers watch and label the same media files
    - Have multiple participants rate their experience of the same tasks
    - Have the same participants engage in the same tasks in different settings
  - These measurements are then compared using statistical methods
    - There are many approaches to estimating the different types of reliability
    - Each approach has its own set of advantages and disadvantages

Inter-Rater Reliability

- Why focus on inter-rater reliability?
  - The methods used for all types of reliability are similar (or identical)
  - The most common use of reliability in AC is between raters for labels
  - This allows you to provide evidence that your labels are reliable/valid
  - When there is no ground truth, we settle for consistency among raters
- What specific approaches will we explore?
  - For categorical measurements, we will discuss agreement indexes
  - For dimensional measurements, we will discuss correlation coefficients

Agreement: Two raters and two categories

Object	Rater 1	Rater 2	Match?	Observed/Possible
1	1	1	Agree	1/1
2	0	1	Disagree	0/1
3	0	0	Agree	1/1
4	1	0	Disagree	0/1
5	1	1	Agree	1/1
6	0	1	Disagree	0/1
7	1	1	Agree	1/1
8	0	0	Agree	1/1
9	0	0	Agree	1/1
10	1	1	Agree	1/1

{0, 1} = {No Smile, Smile}

$M = .70$

Agreement: Many raters and categories

	R1	R2	R3	R1-R2	R1-R3	R2-R3	O/P
1	2	2	2	Agree	Agree	Agree	3/3
2	1	3	1	Disagree	Agree	Disagree	1/3
3	3	3	3	Agree	Agree	Agree	3/3
4	1	2	2	Disagree	Disagree	Agree	1/3
5	1	1	1	Agree	Agree	Agree	3/3
6	1	2	2	Disagree	Disagree	Agree	1/3
7	1	1	1	Agree	Agree	Agree	3/3
8	3	3	3	Agree	Agree	Agree	3/3
9	2	2	3	Agree	Disagree	Disagree	1/3
10	2	2	2	Agree	Agree	Agree	3/3
{1, 2, 3} = {Math, Science, Art}				$M = .70$	$M = .70$	$M = .80$	$M = .73$

Agreement: Ordered categories

	R1	R2	R3
1	2	2	2
2	1	3	1
3	3	3	3
4	1	2	2
5	1	1	1
6	1	2	2
7	1	1	1
8	3	3	3
9	2	2	3
10	2	2	2

{1, 2, 3} = {Low, Medium, High}

- Weighting schemes for ordered categories
  - Identity = Same as before (unordered)
  - Linear = Credit is equally spaced
  - Quadratic = Credit decays

Credit Awarded with 3 Categories

	Same	1 Away	2 Away
Identity	1.00	0.00	0.00
Linear	1.00	0.50	0.00
Quadratic	1.00	0.75	0.00

Agreement: Ordered categories

	R1	R2	R3	R1-R2	R1-R3	R2-R3	O/P
1	2	2	2	1.00	1.00	1.00	3/3
2	1	3	1	0.00	1.00	0.00	1/3
3	3	3	3	1.00	1.00	1.00	3/3
4	1	2	2	0.50	0.50	1.00	2/3
5	1	1	1	1.00	1.00	1.00	3/3
6	1	2	2	0.50	0.50	1.00	2/3
7	1	1	1	1.00	1.00	1.00	3/3
8	3	3	3	1.00	1.00	1.00	3/3
9	2	2	3	1.00	0.50	0.50	2/3
10	2	2	2	1.00	1.00	1.00	3/3

{1, 2, 3} = {Low, Medium, High}

$M = .83$

Agreement: Generalized function

```
% Import data from CSV file
>> CODES = csvread('Categorical-Data.csv');

%Compute agreement for unordered categories
>> mAGREE(CODES, 1:3, 'identity')
Percent observed agreement = 0.675

% Compute agreement for linear categories
>> mAGREE(CODES, 1:3, 'linear')
Percent observed agreement = 0.838

% Compute agreement for quadratic categories
>> mAGREE(CODES, 1:3, 'quadratic')
Percent observed agreement = 0.919
```

Example Dataset

	R1	R2	R3	R4	R5
1	2	2	3	2	2
2	2	2	2	2	2
3	2	NaN	2	2	1
4	1	2	2	2	2

Amount of Credit Awarded

	Same	1 Away	2 Away
Identity	1.00	0.00	0.00
Linear	1.00	0.50	0.00
Quadratic	1.00	0.75	0.00

Agreement: Issues

- Are there issues with agreement?
  - What if raters guess and end up agreeing by chance?
  - What is the right “baseline” to compare agreement to?
  - What if some categories are more common than others?
  - What if agreement is higher for some categories than others?
- Can we address these issues?
  - Chance-adjusted agreement indexes try to address the first two issues
  - Category-specific agreement indexes try to address the last two issues

Agreement: Chance-adjusted

- What is a chance-adjusted agreement index?
  - How much agreement would occur "by chance" alone (p_c)?
  - If we know p_c, we can adjust observed agreement by this amount

$$r_i = \frac{p_o - p_c}{1 - p_c} = \frac{\text{Observed Nonchance Agreement}}{\text{Possible Nonchance Agreement}}$$

  - This yields the general form of a chance-adjusted agreement index (r_i)
  - It is still the ratio of observed to possible agreement, but p_c is removed
  - "When raters could agree *honestly*, how much did they do so?"

Agreement: Chance-adjusted

- How can chance agreement be estimated?
  - We need to build a "baseline" model to compare raters to
  - In practice, p_c is only an estimate of chance agreement
  - Different chance-adjusted indexes are based on different assumptions
  - They usually use the same general form (r_i) but estimate p_c differently
  - There are two primary types of assumptions about chance agreement

Agreement: Chance-adjusted

- What are the category-based assumptions?
  - Each category has an equal probability of being randomly selected
  - So chance is modeled as “flipping coins” or “rolling dice”
  - Bennett et al.’s (1954) S score was the first version of this approach
- What are the distribution-based assumptions?
  - Each category’s probability of being randomly selected is equal to its prevalence
  - So chance is modeled as “meeting a quota” for each category
  - Quotas may be rater-specific (Cohen’s κ) or shared (Scott’s π, Krippendorff’s α)

Agreement: Chance-adjusted

```
% Compute S for unordered categories
>> mSSCORE(CODES, 1:3, 'identity')
Percent observed agreement = 0.675
Percent chance agreement = 0.333
Bennett et al.'s S score = 0.513
```

Bennett et al.'s S score = 0.513

```
% Compute kappa for unordered categories
>> mKAPPA(CODES, 1:3, 'identity')
Percent observed agreement = 0.675
Percent chance agreement = 0.725
Cohen's kappa coefficient = −0.182
```

Cohen's kappa coefficient = −0.182

Example Dataset

	R1	R2	R3	R4	R5
1	2	2	3	2	2
2	2	2	2	2	2
3	2	NaN	2	2	1
4	1	2	2	2	2

Agreement: Category-specific

- What is the category-specific agreement index?
  - What if some categories are more difficult/ambiguous than others?
  - To explore this, we can calculate agreement for specific categories

$$SA_k = \frac{\text{Observed Agreement on Category } k}{\text{Possible Agreement on Category } k}$$

  - SA_k is the conditional probability of a random rater assigning a random object to category k given that another random rater already did so
  - SA_k is based on the work of Dice (1945) and Sørensen (1948)

Agreement: Category-specific

```
% Compute specific agreement
>> mSPECIFIC(CODES, 1:3, 'identity')
Specific agreement for category 1 = 0.000
Specific agreement for category 2 = 0.820
Specific agreement for category 3 = 0.000
```

Example Dataset

	R1	R2	R3	R4	R5
1	2	2	3	2	2
2	2	2	2	2	2
3	2	NaN	2	2	1
4	1	2	2	2	2

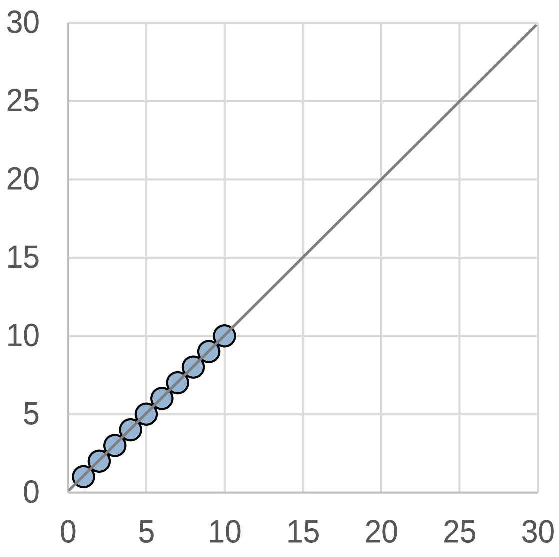
Compare these SA_k results to $S = 0.513$ and $\kappa = -0.182$
Each approach tells a very different story about reliability.
Which assumptions are we most comfortable with?
What are the pros and cons of each approach?

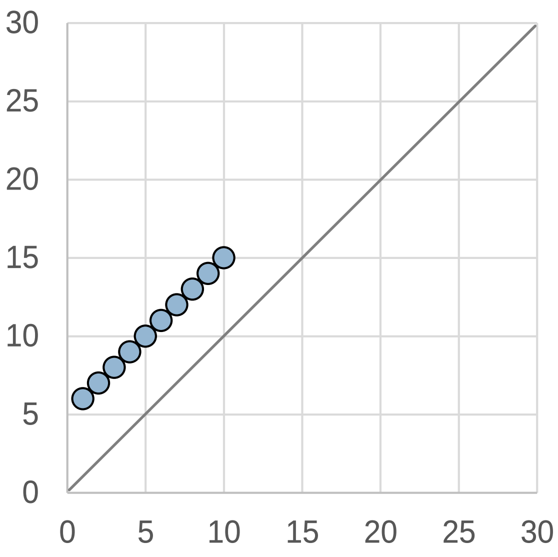
Correlations

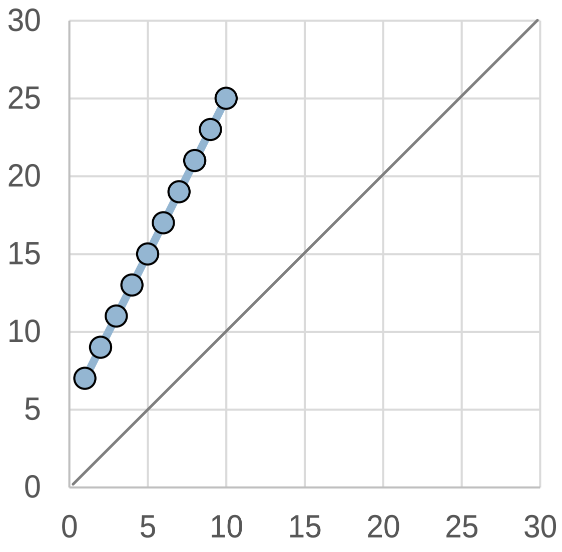
- What is a correlation coefficient?
  - Variance is a measure of the amount of spread or dispersion in a variable
  - Variance comes from difference sources and can be partitioned by source
  - Correlation coefficients are normalized measures of co-variance (-1 to 1)
  - Various correlation coefficients can be used to measure reliability
  - *We will discuss several intra-class correlation coefficients (ICCs)*

Correlations: Agreement and consistency

Agreement ICC	Consistency ICC
Requires $Y = X$	Allows $Y = X + b$







$A = 1.0$	$C = 1.0$	$A = 0.4$	$C = 1.0$	$A = 0.2$	$C = 0.8$

Correlations: Agreement and consistency

Agreement ICC (Intra-class correlation)	**Consistency ICC** (Intra-class correlation)
$A = \frac{\sigma^2_{row}}{\sigma^2_{row} + \sigma^2_{col} + \sigma^2_{err}}$	$C = \frac{\sigma^2_{row}}{\sigma^2_{row} + \sigma^2_{err}}$
$\frac{\text{Object}}{\text{Object} + \text{Rater} + \text{Error}}$	$\frac{\text{Object}}{\text{Object} + \text{Error}}$
High A comes from high object, low rater, and low error variance	High C comes from high object variance and low error variance

Note. Because it is the numerator in the ICC formulas, low object variance makes a high ICC almost impossible.

Correlations: Agreement and consistency

```
% Import data from CSV file
>> RATINGS = csvread('Dimensional-Data.csv');

% Compute ICCs for single measures
>> ICC_C_1(RATINGS)
Single measures consistency ICC = 0.622
>> ICC_A_1(RATINGS)
Single measures agreement ICC = 0.558

% Compute ICCs for average measures
>> ICC_C_k(RATINGS)
Average measures consistency ICC = 0.767
>> ICC_A_k(RATINGS)
Average measures agreement ICC = 0.716
```

Example dimensional data

	R1	R2
1	7.800	7.800
2	7.800	-34.000
3	42.170	-120.556
4	101.950	-123.600
5	184.033	-151.630
…	…	…

Inter-Rater Reliability

- Where can I read more?
  - Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Gaithersburg, MD: Advanced Analytics.
  - McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.
  - Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind inter-coder reliability indices. In C. T. Salmon (Ed.), *Communication Yearbook* (pp. 418–480). Routledge.
- Where can I find those functions?
  - http://mreliability.jmgirard.com (MATLAB)
  - http://www.agreestat.com/r_functions.html (R)