Assessing inter-rater agreement for nominal judgement variables

Gareth McCray

Please Cite as:

McCray, G, (2013) *Assessing inter-rater agreement for nominal judgement variables.* Paper presented at the Language Testing Forum. Nottingham, November 15-17.

Definition of terms

Inter-rater agreement	Nominal variable	Judgement variable
In the context of this talk it is defined as *the extent to which raters agree with one another* Inter-rater agreement might also be termed "inter-rater reliability" or "rater concordance"	Variables which cannot be measured on a scale and have no implied hierarchical relationship to one another E.g. - What is the first language of the student? - Did the student exhibit a specific behaviour (yes/no)?	Variables which reflect the subjective, yet informed opinion of a judge about a specific matter under investigation

Background to the study

- This study is part of an initial phase of a project attempting to model item difficulty in reading comprehension in the L2 context, which made use of expert judgments.

- Before statistical modelling could take place, the levels of inter-rater agreement had to be evaluated in order to verify that only variables with statistically significant levels of agreement were included in the model.

- Judgements were gathered on 8 **dichotomous** variables on 82 items.

Communicating inter-rater agreement

"However, the reliability of the expert judges was rather low …"

"The overall level of agreement was very high…"

"Compared with what?"

"According to which benchmark?"

Statement of the Problem (1)

- Various methods have been used in language testing to communicate a summary value of inter-rater agreement.
- This variety of methods makes it difficult to make comparisons between different studies.
- To allow the levels of agreement of expert judgements to be compared across studies it would be a good idea to have a robust standard statistic for reporting levels of agreement which has published benchmark levels.

Method	Problems…
Summary Statistics Raw Agreement Proportion/Percentage	No measure of statistical significance. No inclusion of agreement due to chance.
Cronbach's Alpha	A measure of internal consistency not of rater agreement *per se.*
G-Theory variance components	Difficult to communicate the results.

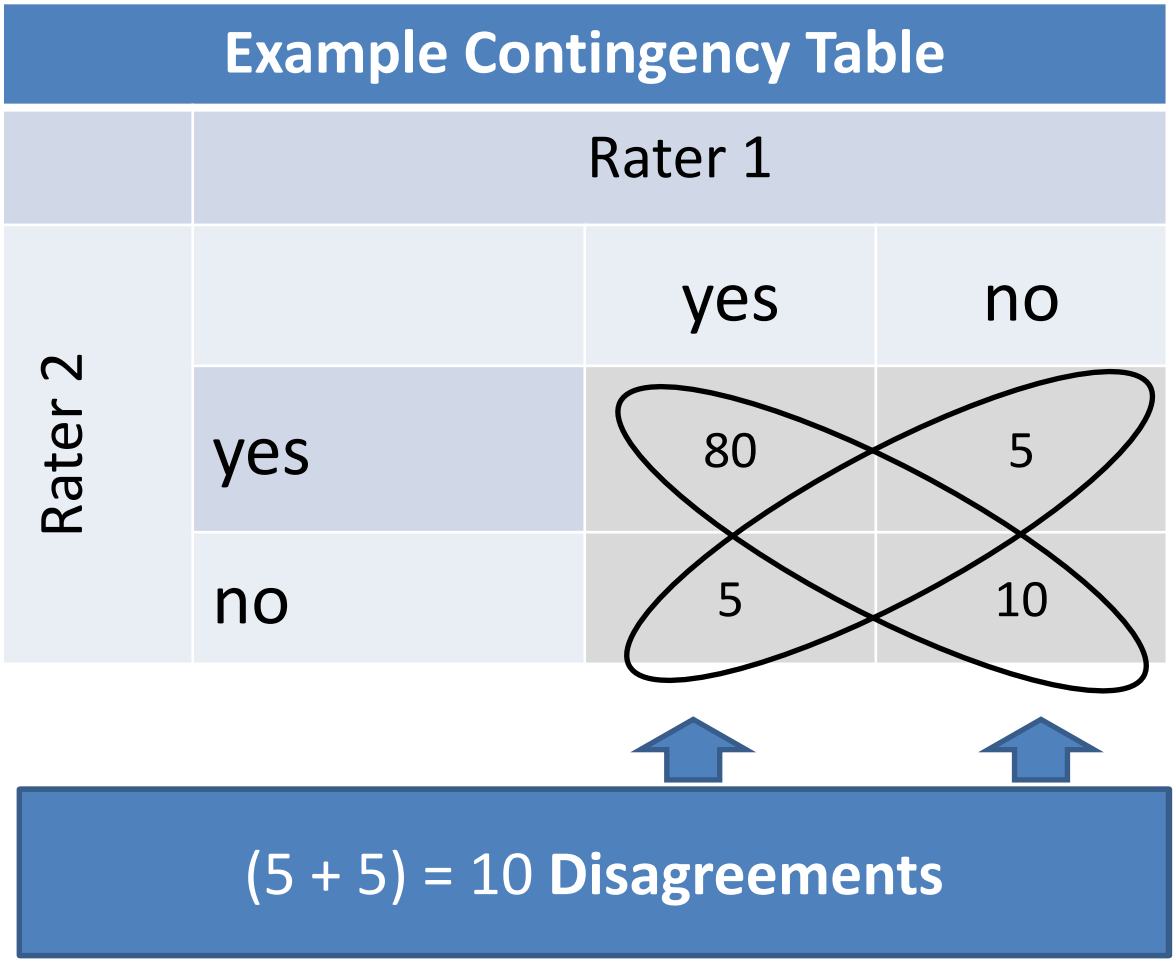
Statement of the Problem (2)

- The fact that the data were dichotomous and that many had ***'high trait prevalence'*** meant that a suitable technique for the assessment of inter-rater agreement was problematic.

- I intend to present some methods for the assessment of inter-rater agreement for this data and discuss their advantages and disadvantages.

Evaluative criteria	
Interpretability & Communicability	✔
Robustness	✔

Contingency Tables



Raw Agreement Proportion (RAP)

Raw Agreement Proportion			
	Coin 1		
Coin 2		Heads	Tails
	Heads	25	22
	Tails	23	30

RAP = 55%

Problems with the RAP

No consideration of agreement due to chance

No test of statistical significance

If we have two raters rating on a dichotomous scale **at random** there is a 50% chance of agreement.

If we have 2 raters rating on a 3 point scale (low/medium/high) **at random** there is a 33% change of agreement.

The RAP does not allow us to (simply) measure perform a test of whether the agreements found are statistically significantly better than those that can be attributed to chance alone.

Evaluative criteria	
Interpretability & Communicability	X
Robustness	

Cohen's Kappa

- Cohen's Kappa (1960), one of the most widely used indices of rater agreement, overcomes this problem of chance agreement.
- It gives a value of between 0 and 1; 0 being the lowest level of agreement 1 being the highest.
- Clear benchmarks for the strength of agreement have been constructed for its use to aid in communicability.

Fleiss' (1981) Benchmark Scale for the Kappa	
< 0.40	Poor
0.40 to 0.75	Intermediate to good
More than 0.75	Excellent

Landis and Koch (1977) Benchmark Scale for the Kappa	
< 0.0	Poor
0.00 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost perfect

Altman's (1991) Benchmark Scale for the Kappa	
< 0.20	Poor
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Good
0.81 to 1.00	Very good

Sensitivity to ***Marginal Homogeneity***

Example contingency table 1

		Rater 1	
		yes	no
Rater 2	yes	30	20
	no	20	30

RAP = 60%

Cohen's Kappa = 0.200*
Krippendorff's Alpha = 0.204*
Phi = 0.200*

Example contingency table 2

		Rater 1	
		yes	no
Rater 2	yes	30	30
	no	10	30

RAP = 60%

Cohen's Kappa = 0.231*
Krippendorff's Alpha = 0.204*
Phi = 0.250*

Sensitivity to ***Trait Prevalence***

Example contingency table 3

		Rater 1	
		yes	no
Rater 2	yes	45	5
	no	5	45

RAP = 90%

Cohen’s Kappa = 0.800***
Krippendorff's Alpha = 0.801***
Phi = 0.800***

Example contingency table 4

		Rater 1	
		yes	no
Rater 2	yes	80	5
	no	5	10

RAP = 90%

Cohen’s Kappa = 0.608***
Krippendorff's Alpha = 0.610***
Phi = 0.608***

Sensitivity to ***Trait Prevalence***

Example contingency table 5

		Rater 1	
		yes	no
Rater 2	yes	45	5
	no	5	45

Cohen's Kappa = 0.800***

Fleiss' Benchmark = **Excellent**
Landis and Koch Benchmark = **Substantial**
Altman's Benchmark = **Good**

Example contingency table 6

		Rater 1	
		yes	no
Rater 2	yes	80	5
	no	5	10

Cohen's Kappa = 0.608***

Fleiss' Benchmark = **Intermediate to Good**
Landis and Koch Benchmark = **Moderate**
Altman's Benchmark = **Moderate**

Problems With Cohen's Kappa

- Cohen's Kappa (and the Phi Coefficient) are sensitive to ***marginal homogeneity*** and as a measure of "agreement".

- Cohen's Kappa (the Phi Coefficient and Krippendorff's Alpha) are sensitive and negatively bias for high ***trait prevalence***.

- For more robust measures of inter-rater agreement, particularly in cases of high ***trait prevalence,*** an alternative statistic is required.

Evaluative criteria	
Interpretability & Communicability	✔
Robustness	X

Gwet's AC1

- Gwet's (2008) AC1 is an alternative statistic
- Like the Kappa it takes a value between 0 and 1
- It is recommended that the same benchmark scales as those for the Kappa can be used for communication of levels of agreement.
- It is not sensitive to ***marginal homogeneity*** and positively biases for ***trait prevalence***.
  - It can be extended to multiple raters
  - It can deal with both nominal and ordinal data
  - it can deal with missing data

Evaluative criteria	
Interpretability & Communicability	✔
Robustness	✔

AC1 and ***Marginal Homogeneity***

Example contingency table 7

		Rater 1	
		yes	no
Rater 2	yes	30	20
	no	20	30

RAP = 60%

Cohen's Kappa = 0.200*

Gwet's AC1 = 0.200*

Example contingency table 8

		Rater 1	
		yes	no
Rater 2	yes	30	30
	no	10	30

RAP = 60%

Cohen's Kappa = 0.231*

Gwet's AC1 = 0.200*

AC1 and ***Trait Prevalence***

Example contingency table 9

		Rater 1	
		yes	no
Rater 2	yes	45	5
	no	5	45

RAP = 90%

Cohen's Kappa = 0.800***

Gwet's AC1 = 0.800***

Example contingency table 10

		Rater 1	
		yes	no
Rater 2	yes	80	5
	no	5	10

RAP = 90%

Cohen's Kappa = 0.608***

Gwet's AC1 = 0.866***

AC1 ***Trait Prevalence*** Increase

Example contingency table 11

		Rater 1	
		yes	no
Rater 2	yes	30	20
	no	20	30

RAP = 60%

Cohen's Kappa = 0.2*
Gwet's AC1 = 0.2*

Example contingency table 12

		Rater 1	
		yes	no
Rater 2	yes	50	20
	no	20	10

RAP = 60%

Cohen's Kappa = 0.04
Gwet's AC1 = 0.31**

Substantive Results

Overall measures on inter-rater agreement	
AC1	0.58**
Kappa	0.34**
RAP	0.82

Measures of inter-rater agreement by item type			
	SAMC	MAMC	GAPS
AC1	0.39**	0.52**	0.57**
Kappa	0.39**	0.47**	0.29**
RAP	0.81	0.84	0.79

Measures of inter-rater agreement by variable

	Proximity of pieces of information	*Competing information in text*	*Prominence of necessary information*	*Semantic match between item and text*	*Concreteness of necessary information*	*Familiarity of topic*	*Register of text*	*Extent to which outside information is required*
AC1	0.46**	0.27**	0.47**	0.08	0.41**	0.48**	0.49**	0.60**
Kappa	0.40**	0.27**	0.07	0.15	0.05	0.06	0.00	0.06
RAP	0.84	0.80	0.85	0.75	0.80	0.82	0.82	0.86

Altman's Benchmark

Overall measures on inter-rater agreement	
AC1	Moderate
Kappa	Fair
RAP	0.82

Measures of inter-rater agreement by item type			
	SAMC	MAMC	GAPS
AC1	Fair	Moderate	Moderate
Kappa	Fair	Moderate	Fair
RAP	0.81	0.84	0.79

Measures of inter-rater agreement by variable

	Proximity of pieces of information	*Competing information in text*	*Prominence of necessary information*	*Semantic match between item and text*	*Concreteness of necessary information*	*Familiarity of topic*	*Register of text*	*Extent to which outside information is required*
AC1	**Moderate**	**Fair**	**Moderate**	**Poor**	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Kappa	**Moderate**	**Fair**	**Poor**	**Poor**	**Poor**	**Poor**	**Poor**	**Poor**
RAP	0.84	0.80	0.85	0.75	0.80	0.82	0.82	0.86

Conclusion

- In order to allow the levels of agreement of expert judgements to be compared across different studies it would be a very good idea to have **a standard statistic** for reporting levels of agreement.

- I would suggest the usage of **Gwet's AC1** as it provides an interpretable, communicable and robust method to disseminate levels of inter-rater agreement.

References

- **Gwet, L. (2010). *The Handbook of Inter-Rater Reliability*. Gaithersburg: Advanced Analytics.**

- Gwet, L. (2008). Computing Inter-Rater Reliability and its Variance in the Presence of High Agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29 – 48.

- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*, 37 – 46.