

Fleiss' Kappa

Verified Python analysis transcript

Test: Fleiss' Kappa

Design: Agreement among three grade classifications (G1, G2, G3) across four categories.

Variables: G1 grade band, G2 grade band, G3 grade band

Null hypothesis: Agreement among the three classifications equals agreement expected from overall category prevalences.

Primary formula: $\text{kappa}_F = (P_{\text{bar}} - P_e) / (1 - P_e)$, $P_i = [\sum_j n_{ij}^2 - n] / [n(n-1)]$

Valid source rows: 649

Verified results:

fleiss_kappa: 0.6118371867294158

observed_agreement: 0.7575757575757576

expected_agreement: 0.37545732322572833

subjects: 649

raters: 3

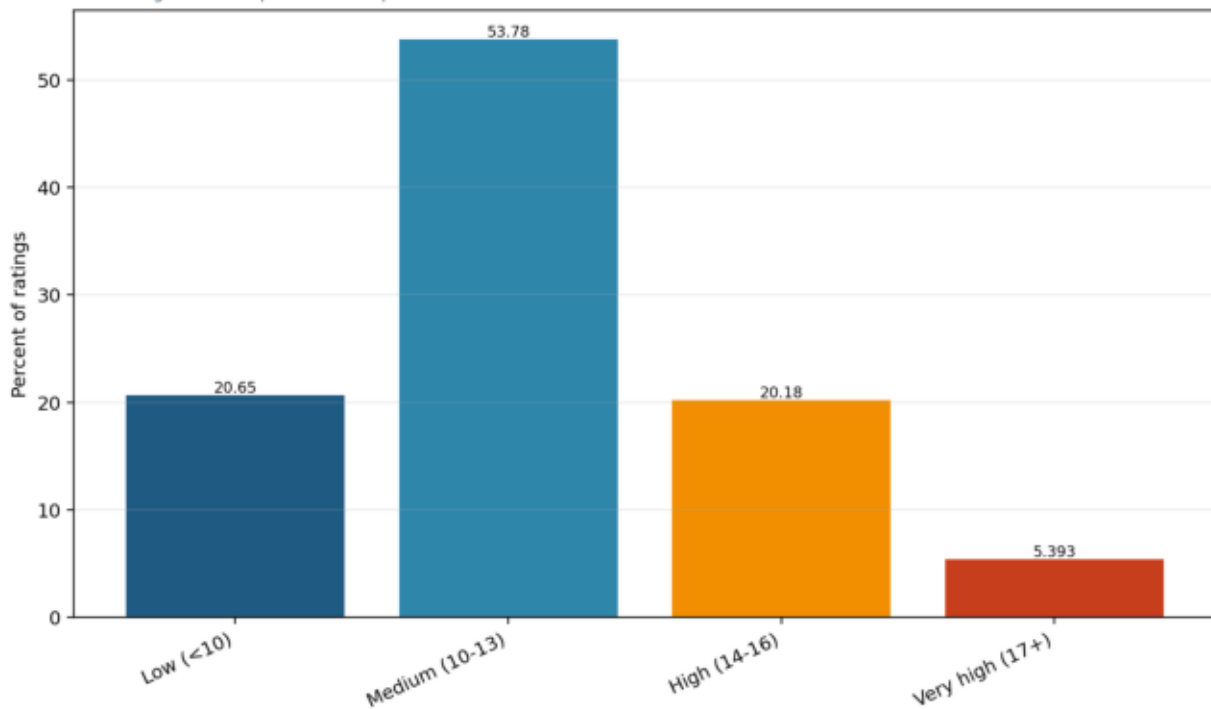
Interpretation:

Fleiss kappa = 0.611837; observed agreement = 0.757576, expected agreement = 0.375457.

This is a multi-rater calculation across G1, G2, and G3, not an average of pairwise Cohen kappas.

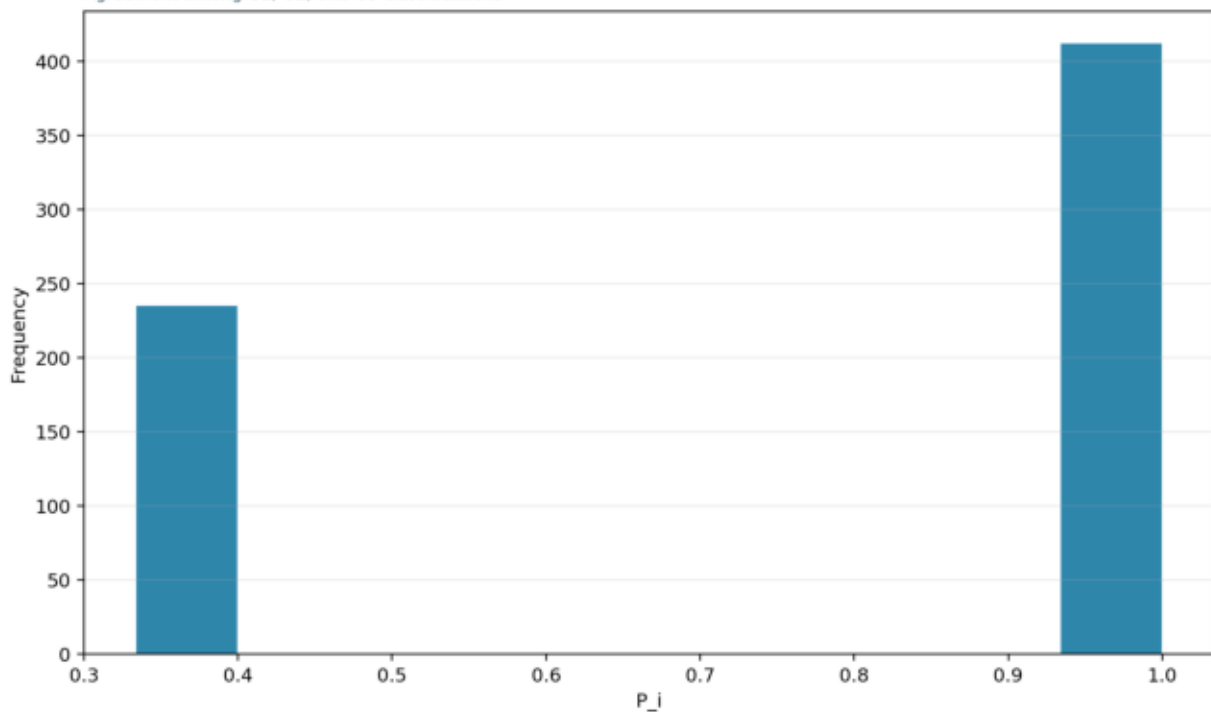
Category prevalence across three grade occasions

Chance agreement is prevalence-dependent



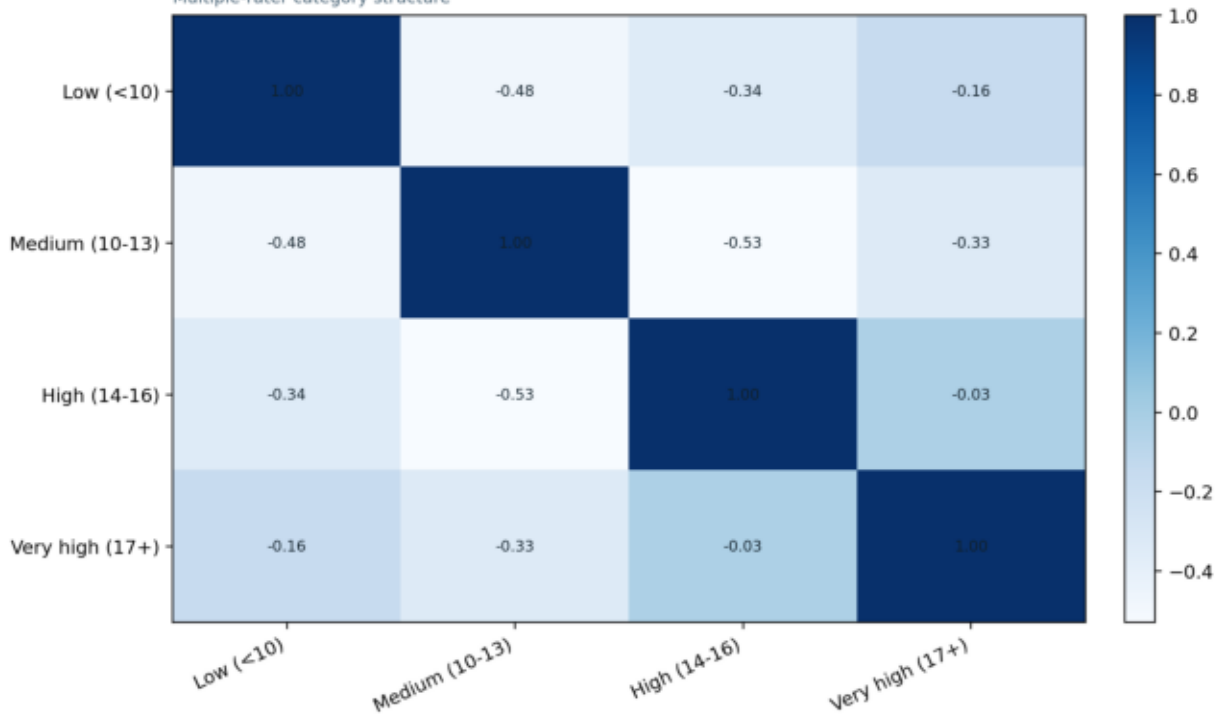
Subject-level inter-rater agreement

Agreement among G1, G2, and G3 classifications



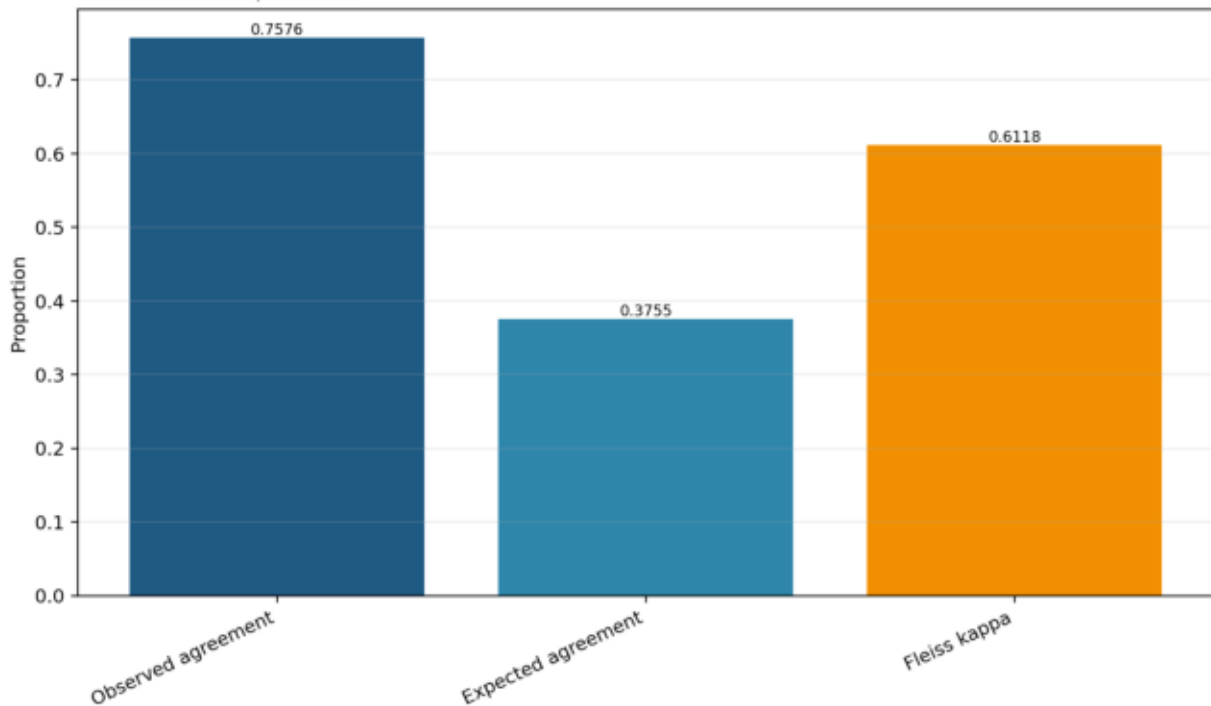
Category-count relationship across subjects

Multiple-rater category structure



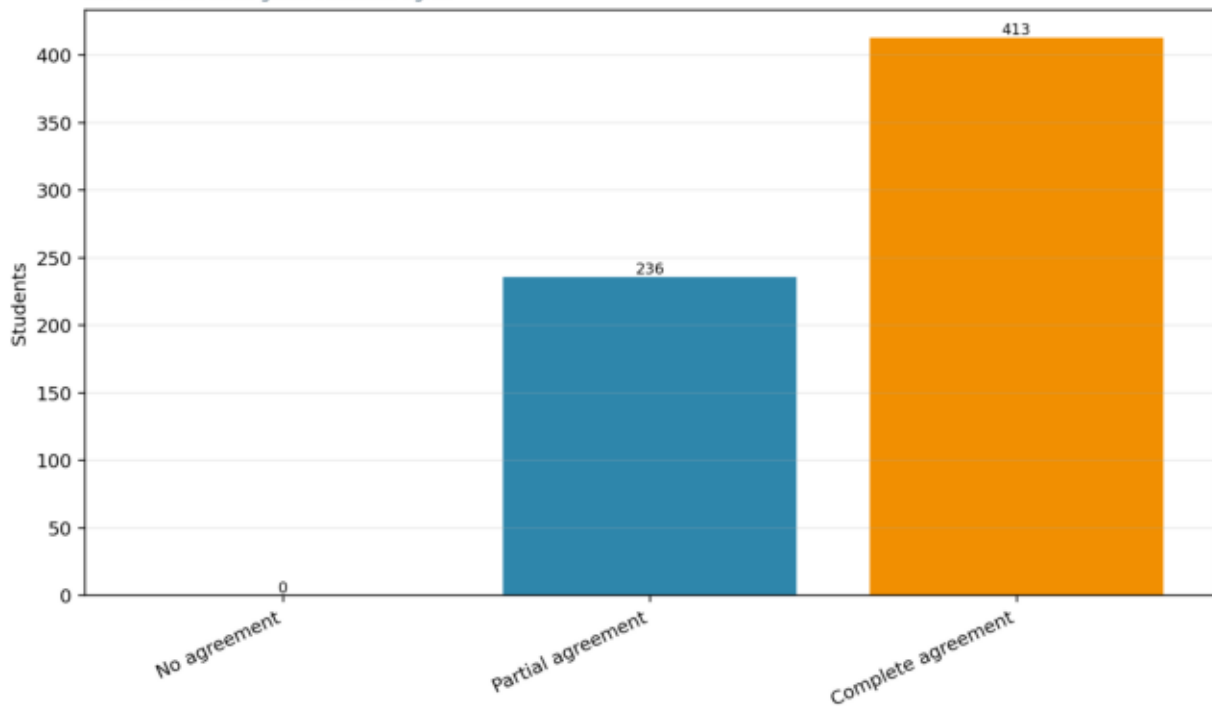
Fleiss kappa components

Three classifications per student



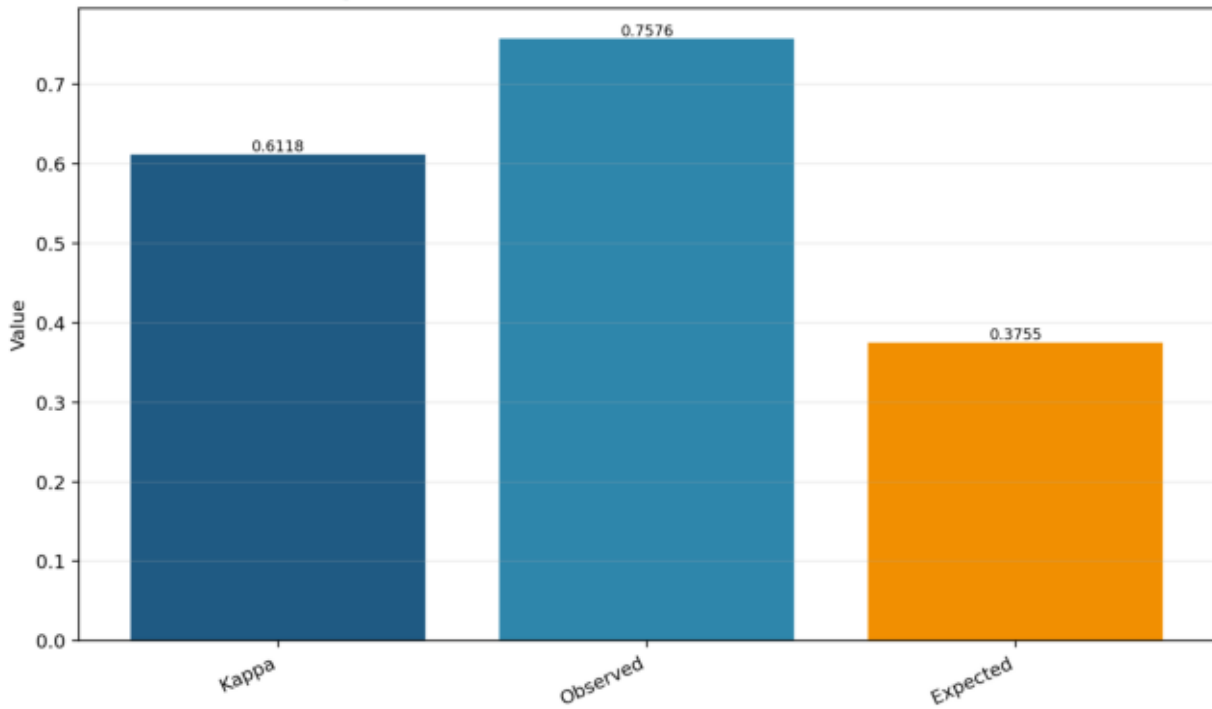
Student agreement-pattern counts

How often the three grade occasions agree



Fleiss kappa result

Chance-corrected multi-rater agreement



Subject Rating Counts

Low (<10)	Medium (10-13)	High (14-16)	Very high (17+)
1	2	0	0
1	2	0	0
0	3	0	0
0	0	3	0
0	3	0	0
0	3	0	0
0	3	0	0
0	3	0	0
0	0	2	1
0	3	0	0
0	0	3	0
0	3	0	0
0	3	0	0
0	3	0	0
0	0	3	0
0	0	0	3
0	2	1	0
0	1	2	0
3	0	0	0
0	3	0	0
0	2	1	0
0	3	0	0
0	2	1	0
0	3	0	0
0	3	0	0
0	3	0	0
0	3	0	0
0	3	0	0

Category Proportions

proportion
0.206471
0.53775
0.201849
0.0539291

Item Agreement

agreement
0.333333
0.333333
1
1
1
1
1
1
1
0.333333
1
1
1
1
1
1
1
0.333333
0.333333
1
1
0.333333
1
0.333333
1
1
1
1
1