

Fleiss Multi-Rater Kappa

Standalone verified Python analysis

Test: Fleiss Multi-Rater Kappa

Design: Agreement among G1, G2, and G3 treated as three exchangeable raters of three grade categories.

Variables: categorized G1, categorized G2, categorized G3

Null hypothesis: Average observed multi-rater agreement equals chance agreement from pooled category margins.

Formula: $\kappa_F = (P_{\text{bar}} - P_e) / (1 - P_e)$, $P_i = [\sum_j n_{ij}^2 - k] / [k(k-1)]$.

Source rows: 649

Calculated metrics:

fleiss_kappa: 0.6421256617091001

observed_agreement: 0.7842835130970724

expected_agreement: 0.3972284016419714

raters: 3

subjects: 649

Independent reference values:

fleiss_kappa: 0.6421256617091001

observed_agreement: 0.7842835130970724

expected_agreement: 0.3972284016419714

raters: 3

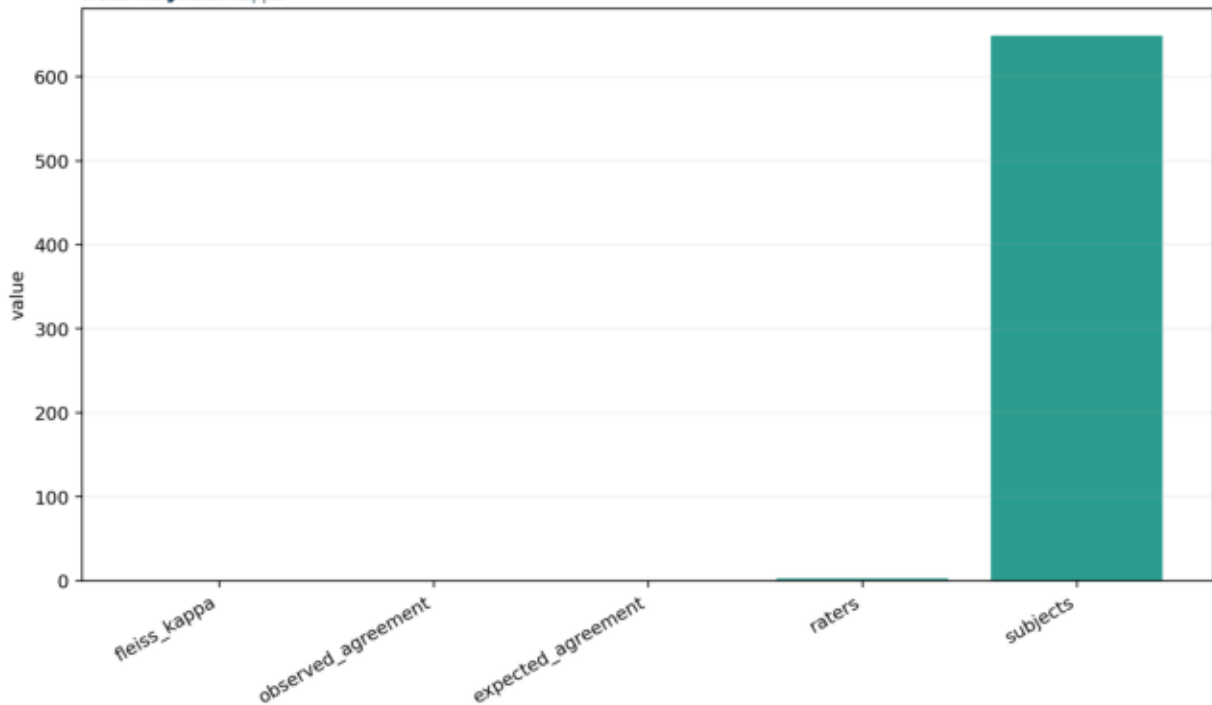
subjects: 649

Interpretation:

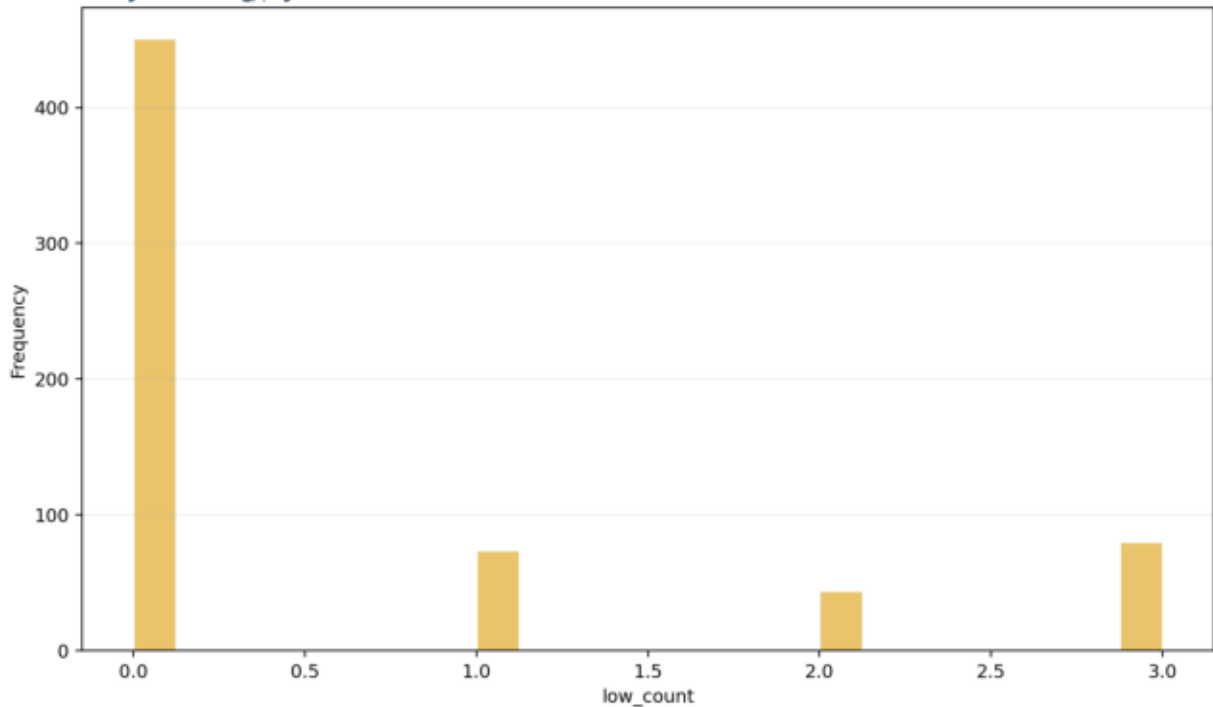
Pooled category margins define chance agreement for all three raters.

Fleiss kappa=0.642126.

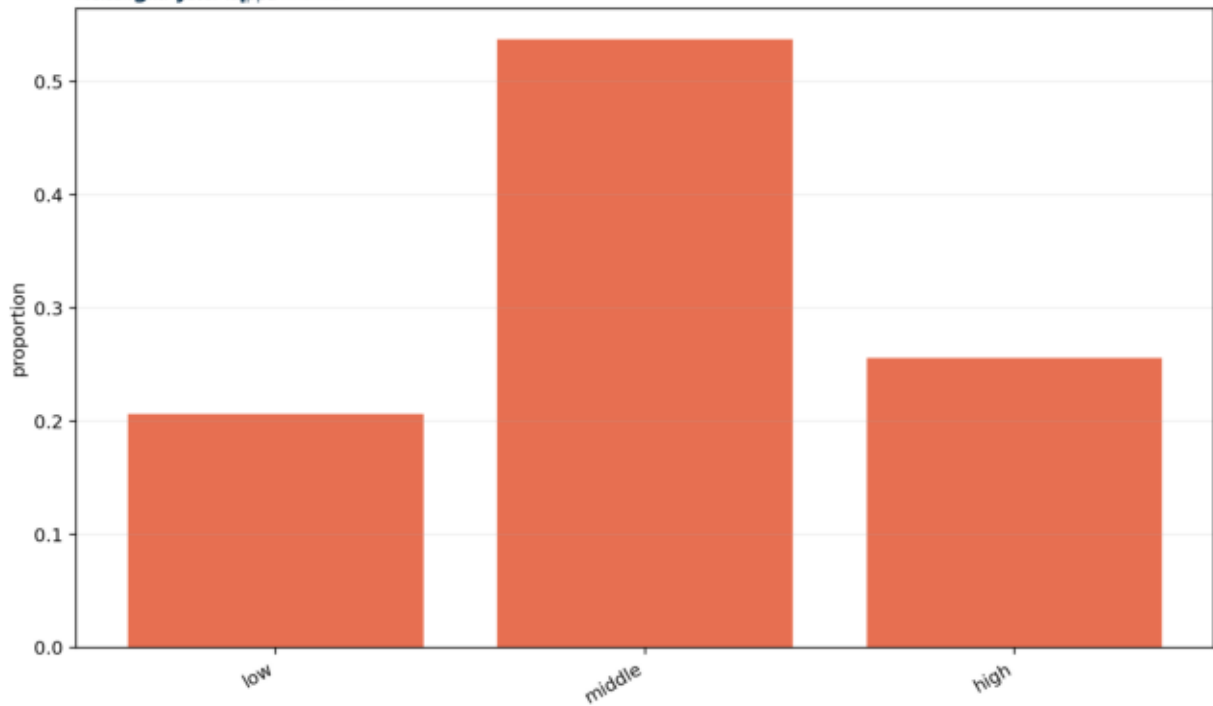
Primary Metrics



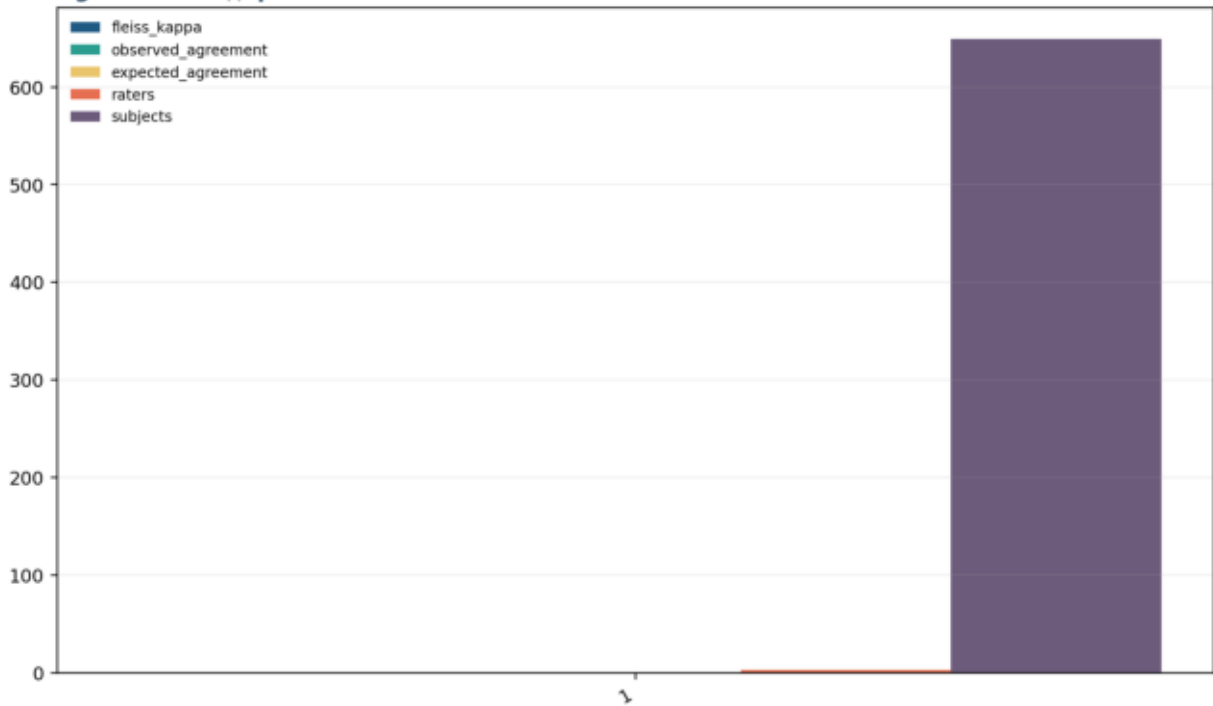
Subject Category Counts



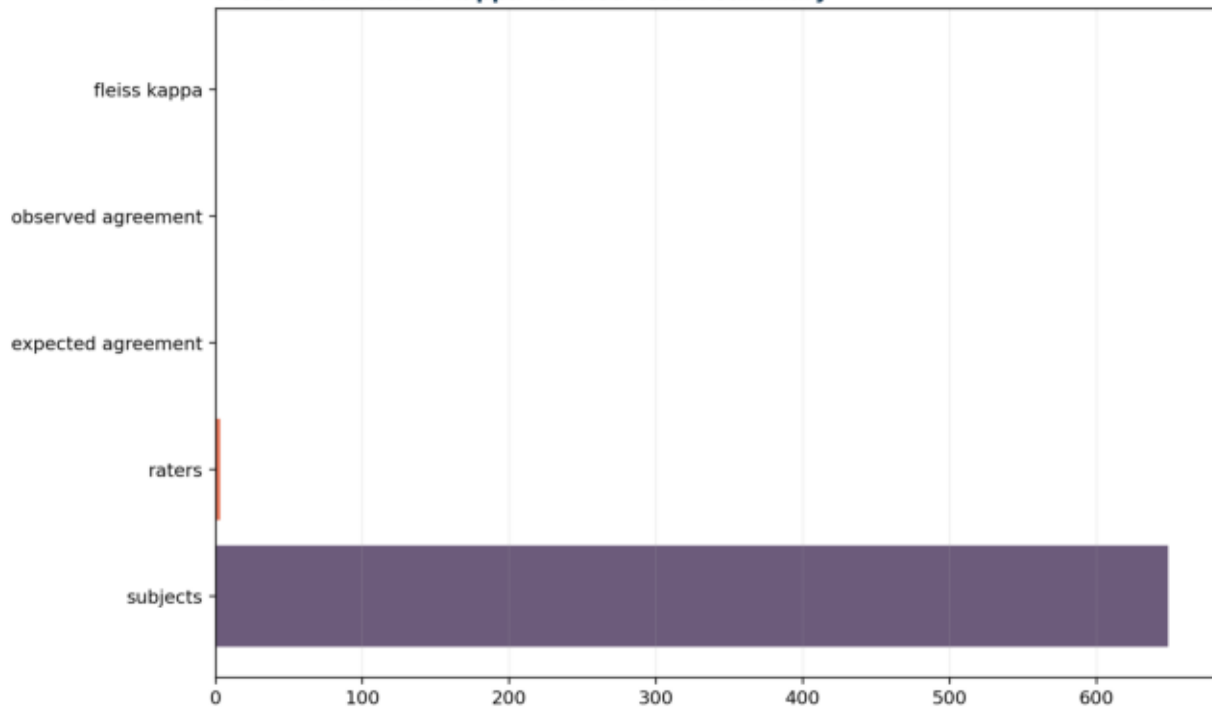
Category Proportions



Agreement Components



Fleiss Multi-Rater Kappa verified result summary



Primary Metrics

metric	value
fleiss_kappa	0.642126
observed_agreement	0.784284
expected_agreement	0.397228
raters	3
subjects	649

Subject Category Counts

low_count	middle_count	high_count	subject_agreement
1	2	0	0.333333
1	2	0	0.333333
0	3	0	1
0	0	3	1
0	3	0	1
0	3	0	1
0	3	0	1
0	3	0	1
0	0	3	1
0	3	0	1
0	0	3	1
0	3	0	1
0	3	0	1
0	3	0	1
0	0	3	1
0	0	3	1
0	2	1	0.333333
0	1	2	0.333333
3	0	0	1
0	3	0	1
0	2	1	0.333333
0	3	0	1
0	2	1	0.333333
0	3	0	1
0	3	0	1
0	3	0	1

Category Proportions

category	proportion
low	0.206471
middle	0.53775
high	0.255778

Agreement Components

fleiss_kappa	observed_agreement	expected_agreement	raters	subjects
0.642126	0.784284	0.397228	3	649