Intro to Intercoder Reliability

Intercoder or inter-rater reliability refers to the degree of agreement among independent coders in their categorization or interpretation of data. High reliability reflects not only consistent application of coding criteria but also a meaningful level of consensus among coders. This suggests that the analysis is not merely subjective, but systematic and replicable. Such consistency and shared understanding are essential for establishing the trustworthiness, rigor, and credibility of research findings.

How is it measured?

There are several ways to measure intercoder reliability, with Cohen's Kappa coefficient (κ) being among the most popular for assessing the agreement between two raters coding nominal data while accounting for chance agreement.

$$k = \frac{p_o - p_e}{1 - p_e}$$

where,

p_o : observed agreements

p_e : expected agreements by chance

How to interpret results?

k	Agreement Level*
0.41 – 0.60	Fair
0.61 – 0.80	Moderate
> 0.81	Strong

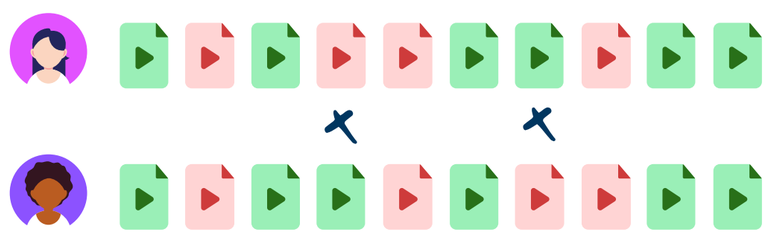
*More conservative thresholds have been proposed, with values below 0.60 often considered indicative of inadequate agreement.

A Strong *k* demonstrates:

- The coding system is clear and transferable, allowing others to understand and apply it in similar contexts.
- Codes are applied consistently across researchers, reducing subjectivity.
- Findings are grounded in data, not personal bias; thereby upporting credibility and confirmability.
- Procedures are clear and documented, ensuring dependability and transparency.

Example

In a study on nonverbal emotional cues, Ruby and Ella were asked to independently watch and rate the same 10 deposition clips, specifically assessing whether the speakers displayed facial expressions of confusion/uncertainty in their testimonies.



Because they disagreed on two videos, the observed agreement is:

$$p_o = \frac{8}{10} = 0.80$$

Ruby and Ella might agree by chance alone. In this example, each has a 50% chance of independently rating a clip as 'yes' or 'no.' Thus, the Cohen's Kappa for this case would be:

$$k = \frac{0.80 - 0.50}{1 - 0.50} = \frac{0.30}{0.50} = 0.60$$

This result may be seen as moderate by lenient standards but inadequate by stricter ones, so it's crucial to justify your chosen threshold within your research context.

There are other methods, such as Krippendorff's Alpha and Fleiss' Kappa, that may be better suited depending on the coding task, the number of coders, the data type, or the presence of missing data. Check the paper below for a comparison:

Halpin, S. N. (2024). Inter-coder agreement in qualitative coding: Considerations for its use. American Journal of Qualitative Research, 8(3), 23-43. https://doi.org/10.29333/ajqr/14887

Did you know that many data analysis tools, including those for qualitative data, have built-in features for checking intercoder reliability? Want to learn more?

rds@library.ucsb.edu