Inter-Rater Reliability Analysis

Inter-rater reliability (IRR) quantifies the extent to which different judges or raters agree in their assessment decisions and is crucial for validating the consistency and reliability of judgments made in survey-based research.

In this study, IRR was assessed using Krippendorff's alpha, a statistical measure that accounts for agreement among raters, both on nominal and ordinal scales. For context-dependent judgments that were not strictly binary, additional measures such as Cohen’s kappa and Intraclass Correlation Coefficients (ICC) were considered to accommodate more nuanced interpretations.

The data were segmented and analyzed by 'Language', 'Aspect' (Context, Culture, Emotion), and 'Type' (Nominal, Ordinal). The overall mean Krippendorff's alpha across all raters was -0.175, with a standard deviation of 0.157. This indicates a low level of agreement, as positive values closer to 1 represent better reliability, and values closer to or below zero suggest no agreement beyond chance.

The analyses were further detailed as follows:

English (EN) showed a higher alpha score in every category compared to Chinese (CN), suggesting more consistent ratings among English-speaking raters.

By aspect, 'Emotion' showed slightly better agreement scores compared to 'Context' and 'Culture', indicating that raters found emotional content somewhat easier to agree upon.

Table 1: Summary of Krippendorff's Alpha by Group

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Language | Aspect | Type | Mean | Min | Max | SD |
| EN | Emotion | Nominal | -0.075 | -0.105 | -0.010 | 0.044 |
| EN | Emotion | Ordinal | -0.107 | -0.197 | -0.022 | 0.073 |
| CN | Emotion | Nominal | -0.153 | -0.664 | 0.082 | 0.187 |
| CN | Emotion | Ordinal | -0.203 | -0.786 | 0.070 | 0.223 |

The low IRR, especially in ordinal judgments and among Chinese raters, may point to cultural differences in interpretation or ambiguity in survey questions. These factors could contribute to the variance observed in initial analyses, as noted by the wide interquartile ranges (IQR) in the box and whisker plots provided. Improving question clarity and considering cultural nuances might help enhance rater agreement in future surveys.

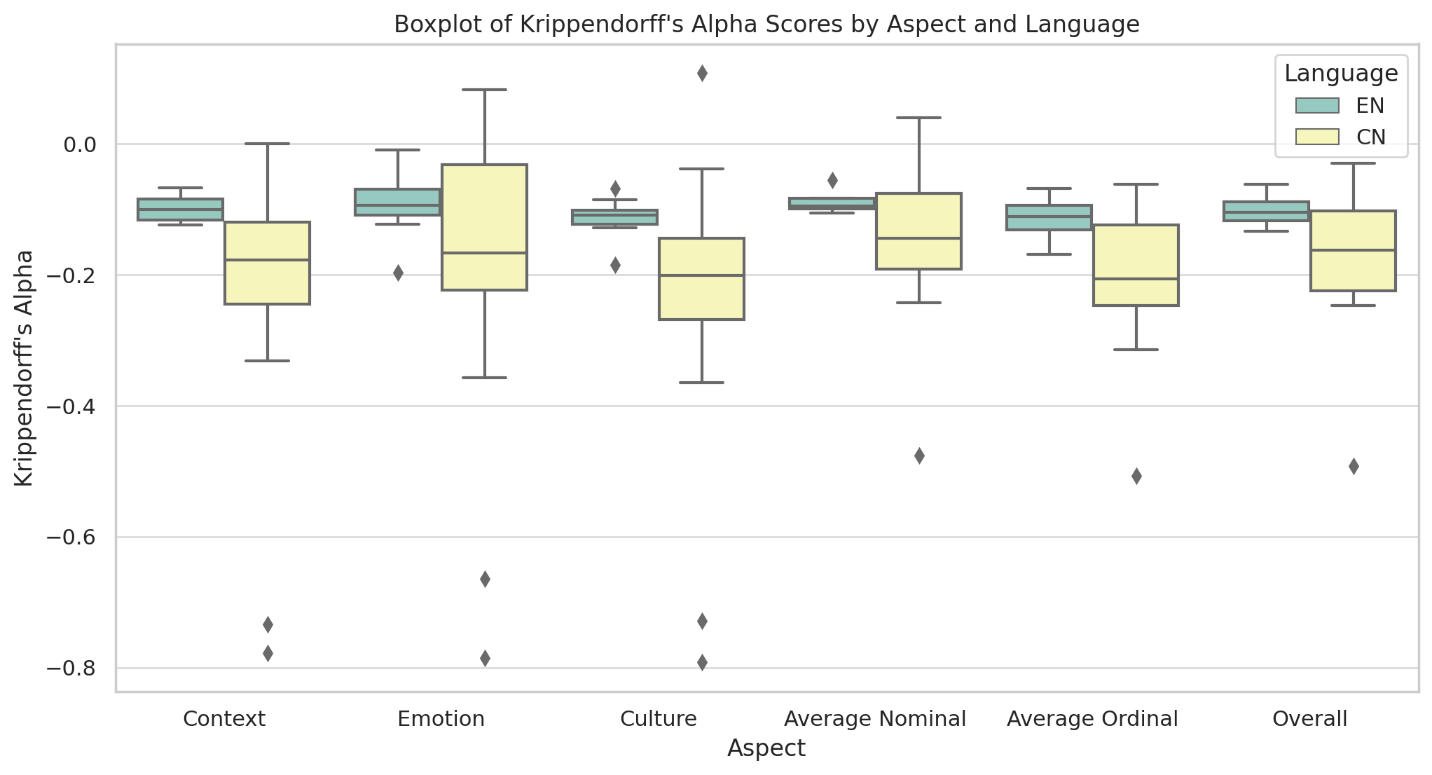


Figure 1. Krippendorff's alpha scores by Aspect Boxplot

In our inter-rater reliability analysis by aspect, the average Krippendorff's alpha scores reveal differing levels of agreement: Context scores averaged -0.203, indicating low agreement, while Culture scored slightly higher at -0.174, and Emotion exhibited a relatively better consensus among raters with an average score of -0.164. Overall, the combined aspects yielded an average agreement score of -0.181. When disaggregated by type, Nominal assessments showed higher agreement levels with an average score of -0.151, compared to Ordinal types, which demonstrated more variability and lower agreement at an average of -0.192. These results underscore the challenges in achieving consistent ratings across different aspects and highlight the greater difficulties encountered with ordinal judgments, suggesting a need for more refined scales or clearer guidelines in survey design to improve reliability.