COMPOSITE INTERRATER RELIABILITY INDEX (CIRI): A NEW INDEX FOR CONSISTENCY IN HIGH VOLUME RATER ENVIRONMENTS

Alper Şahin[1]

Abstract

Introduction

Interrater reliability (IRR) is a widely used metric in educational assessment, measuring the consistency of ratings assigned by multiple raters. IRR indices such as Cohen's Kappa, Pearson correlation, principal component analysis, and many-facets Rasch Model are widely used by educational researchers to measure the degree of agreement among raters (Stemler, 2004).

Problem

At Intensive English programs, it is common to have more than 20-30 raters scoring the same writing or speaking performance tasks. Most commonly used IRR indices ideally work well in contexts with a limited number of raters. However, they become increasingly impractical as the number of raters increases. Moreover, they necessitate pairwise comparisons, model data fit, extensive computation, and specialized software. This requirement poses a serious problem as the number of raters increases limiting their utility and automation in most large-scale real-world scenarios at Intensive English Programs.

To address these challenges, CIRI, a new more practical IRR measure for large-scale assessment scenarios, has been developed. CIRI simplifies the process of calculating interrater reliability by focusing on the total score rather than the individual sub-scores on the rubrics. In this way, CIRI bypasses the inherent difficulty in achieving consensus on individual sub-scores and focuses on the agreement in total score among the raters instead.

Methodology

To calculate CIRI in this sample study, initially, a selection of student performances including three essays, three paragraphs, and three speaking performances were scored by 92 raters. Subsequently, the Estimated True Scores (TE) described by Sahin, (2021) were calculated for these performances. Then, the CIRI scores were calculated at 2.5%, 5%, 7.5%, and 10% tolerance levels. These tolerance levels correspond to the total number of raters divided by the number of raters whose total scores fall within ±2.5%, ±5%, ±7.5%, and ±10% of the highest possible score obtainable from each performance. Finally, the CIRI overall score was obtained by averaging the scores from these tolerance levels for each performance task. The magnitude of the interrater reliability was determined according to the CIRI decision table.

1 Atılım University, alpersahin2@yahoo.com

Findings

The average CIRI scores for speaking tasks indicated a Fair IRR performance, with .16, .33, .55, .62, and .42 at the CIRI ±2.5%, ±5%, ±7.5%, ±10% tolerance levels, and at the CIRI Overall (n 3, j 92) respectively. Similarly, average CIRI scores for paragraph tasks showed a Fair IRR performance with scores of .19, .36, .53, .62, and .43 at the CIRI ±2.5%, ±5%, ±7.5%, ±10% CIRI tolerance levels, and at the CIRI Overall (n 3, j 92) respectively. Lastly, the average CIRI scores for essay tasks also indicated a Fair IRR performance with scores of .22, .46, .66, .74, and .52 at the CIRI ±2.5%, ±5%, ±7.5%, ±10% CIRI tolerance levels, and at the CIRI Overall (n 3, j 92) respectively

Conclusion

This sample study for CIRI suggested that the magnitude of IRR for the paragraph, essay, and speaking tasks was consistently rated as Fair. Additionally, the CIRI Overall score was somewhat higher for essays with scores closer to the limits of the Good category.

Keywords: *Interrater Reliability, Assessing Writing, Assessing Speaking.*

References

Stemler, S. E., (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation 9(1), 1-11. https://doi.org/10.7275/96jp-xz07

Sahin, A. (2021) The rater performance categorization system (RPCS) for intensive English programs. Shanlax International Journal of Education, 9 (3), 225-241. https://doi.org/10.34293/education.v9i3.3986