The 15th International Congress on Mathematical Education
Sydney, 7-14 July, 2024

ESTIMATING INTER-OBSERVER AGREEMENT ON QUALITATIVE DATA WITH A COMPLEX CODING SCHEME

Andrew Baas
Texas State University

Jennifer A. Czocher
Texas State University

Alex White
Texas State University

In this paper, we treat a problem of approximating inter-observer agreement (IOA) for a specific kind of qualitative data, such as videos, where a complex codebook calls for observer-defined analytic units. We lay out criteria that an inter-observer agreement method should satisfy, and we discuss two methods the literature recommends relative to these requirements. While neither method perfectly suits the data, we show that both are promising for quantifying relevant aspects of an observer-defined coding process like relative agreement between coded data.

INTRODUCTION

Qualitative coding is a common method in mathematics education research for analyzing observational data, whether that be classroom recordings, student problem solutions, or individual interviews. Often the coding is performed by multiple researchers trained on a common codebook, special care being taken to ensure consistent application of the codebook. Mixed methods research which uses qualitative coding to analyze observational data is dependent on measures of inter-rater reliability (IRR) to assess and make claims about coder consistency and, ultimately, the internal validity of their methods. Unfortunately, estimating IRR is not always straightforward, and the choice of statistical method used must align with the data and research questions (Jansen et al., 2003). We are interested in a related construct, inter-observer agreement (IOA), which foregrounds agreement between observers as a means for ascertaining the quality of the data, instead of focusing on reliability of an instrument to be used in statistical analysis (Walter et al., 2019). We are particularly concerned with estimating IOA for a specific kind of data common in mathematics education, namely qualitatively coded recordings of teacher-student interactions in classroom or interview settings. In this case, observers using traditional methods will independently define the start and stop times of codable events, resulting in time sequential data (TSD) (Bakeman & Quera, 1995). Methods that are common in statistical analyses (e.g., Cohen's kappa and inter-class correlation (ICC)) are not well-suited to TSD exactly because the analytic units are observer-defined. On the other hand, it is unknown whether methods developed for TSD are suitable for estimating IOA with complex codebooks. Due to the lack of clear available methods, in (Czocher et al., 2024) we developed a set of criteria for evaluating the suitability of IOA methods for interactions data. The primary criterion is how sensitive the method is to what we term *secondary coding errors*. Secondary coding errors (SCEs) occur when observers agree on what code to use but apply the code to (slightly) different intervals. Many IOA methods would count this against agreement, even though the observers substantively agree. Identifying methods which can appropriately calculate IOA on coded data for teacher-student interactions will provide researchers with tools for quantifying and analyzing phenomena like coder drift, assessing how well-trained new coders are, and assessing the internal validity of their data. In this paper, we report on the utility of two methods to approximate IOA for observers: the time-unit kappa (Bakeman et al., 2009) and the

Observer algorithm (Jansen et al., 2003). We chose these two to compare their performance on SCEs because previously they had been compared using simulated data. Our contribution is comparing their utility using real data drawn from interviews.

LITERATURE REVIEW

The defining feature of TSD is that coded events are identified both by their code and by a start and stop time (Bakeman & Quera, 1995). Because the analytic units are not uniform, they do not form a good basis for statistical comparison of coded data. In response, some methods have been developed to quantify agreement by, for example, linking coded events between two observers (EasyDIag in Holle & Rein, 2015), comparing code sequences using string-distance algorithms (GSEQ-DP in Bakeman et al., 2009), and proposing novel measures of similarity (Thomann, 2001). However, TSD presents unique challenges to automated methods because there is a near-zero probability that two independent observers will select the same start and stop times for a codable event. This can lead to three kinds of SCEs: offset errors, splitting errors, and disjoint boundary errors. When an offset error occurs, both observers identify the same event and assign the same code, but they disagree on the exact starting and ending times, yielding overlapping but not identical intervals. The splitting error occurs when observers agree on the code to apply and on the interval it applies to, but disagree on how or whether the coded event should be interposed with (for example) a non-codable or differently-codable event. This can happen with complex codebooks whose codes are not mutually exclusive. The disjoint boundary error occurs when two observers identify the same codable event, agree on the code to apply, but their coded intervals do not overlap. This can happen for codes with a short duration or when a code calls for inferring a behavior. We claim that SCEs are often simply artifacts of observers having to choose time intervals and not indicative of true disagreement or an unstable codebook.

Alongside their sensitivity to SCEs, many IOA methods are also insensitive to errors of omission/commission, i.e. when one observer codes an event which the other entirely fails to code, which we claim are true disagreements. Additionally, some methods result in multiple marked agreements/disagreements per coded event, inflating the perceived number of observer decisions or, in some situations, inflating the agreement or disagreement counts (Holle & Rein, 2015). We take all of these features into account in our evaluation of IOA methods.

We selected two methods for calculating IOA, the time-unit kappa (Bakeman et al., 2009) and Observer IRR algorithm (Jansen et al., 2003), because Bakeman et al. (2009) showed that both methods yielded expected IRR values on simulated data. Both methods were written to calculate IRR, thus we are interested in how they may perform for estimating IOA. For the time-unit kappa method, the video data is broken down into small, regular segments (e.g. every tenth of a second as a unit) and then κ is calculated as if those were the analytic units of the data (Bakeman et al., 2009). The Observer algorithm operates by identifying agreements/disagreements between entire coded events (Jansen et al., 2003). This method avoids breaking single coded events up into potentially hundreds of analytic units by taking multiple passes through the data, using coded event overlap and distance between coded events to determine which ones agree or disagree. Once agreements and disagreements are identified, an agreement matrix is generated and κ is calculated. In both of these methods, Cohen's kappa assumes codebooks are mutually exclusive and exhaustive (MEE), meaning that different codes should never overlap in coded data (mutually exclusive) and they should cover all events of interest (exhaustive). Yet the coded data we consider does not meet the MEE assumption, as coded events can often overlap.

METHODS

We made small modifications to the implementations of the two methods to account for the fact that coded events do overlap in our TSD. Since omission/commission errors introduce a structural error to the similarity matrix, we also used iterative proportional fitting to calculate the κ as suggested by Holle and Rein (2015). The TSD used was collected as part of a study of successful scaffolding moves during cognitive task-based interviews. The data are 52 videos of individual tertiary students working on mathematical modeling tasks during which the interviewer probes the student's mathematical reasoning. The student actions were coded for phases of the modeling cycle (see Blum & Leiß, 2007). The codebook contains 50 subcodes organized into 6 supercodes. It is possible for codes to co-occur. Each video was coded independently by at least two of four trained observers. Pairs of observers changed to mitigate coding drift.

Our evaluation of the selected methods consisted of two complementary analyses: first, how well they aligned with our *a priori* criteria and, second, how well they performed on real interview data. The criteria we used to evaluate the two IOA methods were: (i) whether they require a codebook to satisfy MEE, (ii) how sensitive they are to SCEs, (iii) whether they identify errors of omission/commission, and (iv) whether they preserve a one-to-one correspondence between marked agreements/disagreements with coded events. We additionally identified how feasible it would be to modify these methods to meet these criteria.

Regarding performance, we tested whether each method could reliably distinguish between (A) two observers coding the same video and (B) two observers coding different videos. Additionally, we evaluated the correlation between IOA coefficients generated by both methods. For the first comparison, we created two populations of comparison coefficients for each method. The "false" population ($N = 1200$) contained the IOA coefficients calculated for observers rating distinct videos. The "true" population ($N = 25$) contained the IOA coefficients calculated for observers rating the same video. The false population provided a baseline of IOA coefficients against which the true population could be compared. These populations were compared using Welch's t-test along with the size of the gap between the true and false populations. For correlation, Spearman's rho was calculated between the true populations generated by each method ($N = 51$), answering whether both methods identified the same observer pair/video combinations as having relatively higher (or lower) IOA.

RESULTS

Neither of the IOA methods satisfy all the stated criteria, yet both come close. Since both methods result in a calculation of κ, they inherently make the MEE assumption and so fail that criterion. Since the time-kappa method compares codes second by second, it incorrectly identifies the offset and disjoint boundary errors as disagreements but does correctly identify the splitting error as an agreement if the split is small. The Observer method properly identifies each of the SCEs because it identifies coded events as agreements if (1) they have some overlap or (2) they are within some tolerance of each other. Regarding errors of commission/omission, the time-kappa method does measure these, while the Observer algorithm does not. Finally, neither method preserves a one-to-one mapping between identified agreements or disagreements and coded events.

Regarding the performance of the methods, Welch's t-test suggests that time-unit kappa can distinguish between false ($M = 0.02, SD = 0.02$) and true ($M = 0.41, SD = 0.16$) IOA calculations

($t(24) = 11.9, p < 0.001$). Welch's t-test also suggests that Observer algorithm also can distinguish between false ($M = 0.04, SD = 0.03$) and true ($M = 0.47, SD = 0.16$) IOA calculations ($t(24) = 13.5, p < 0.001$). In both cases, the true and false populations had no overlap. The distance between the largest false coefficient and the smallest true coefficient for time-unit kappa was 0.061 and for the Observer algorithm was 0.094, indicating that the Observer algorithm provided a larger separation between the true and false populations. Spearman's rho calculated on the IOA coefficients of true comparisons ($N = 51$) revealed a strong positive correlation between the time-kappa and Observer IOA coefficients ($r(49) = 0.93, p < 0.001$), suggesting that both methods identified similar relative orderings of the video data from least to greatest IOA.

CONCLUSION

In this paper we described a set of criteria for methods which calculated IOA on coded teacher-student interactions in classroom or interview settings and applied these criteria to two IOA methods for similar data. While neither method satisfied all criteria for TSD, they were both able to identify agreement within our data and provided consistent rankings of observer/video pairings having relatively higher or lower agreements. These results are promising for agreement analysis on coded data of teacher-student interactions. Further research should assess how well the rankings of these methods align with researcher-defined rankings, should investigate more methods, and should determine how existing methods can be adapted to better fit the needs of researchers in this area.

This research was supported by National Science Foundation Grant No. 1750813.

References

Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS & GSEQ*. Cambridge University Press.

Bakeman, R., Quera, V., & Gnisci, A. (2009). Observer agreement for timed-event sequential data: A comparison of time-based and event-based algorithms. *Behavior Research Methods*, *41*(1), 137–147. https://doi.org/10.3758/BRM.41.1.137

Blum, W., & Leiß, D. (2007). How do Students and Teachers Deal with Modelling Problems? In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical Modelling* (pp. 222–231). Woodhead Publishing. https://doi.org/10.1533/9780857099419.5.221

Czocher, J., Baas, A., White, A., & Melhuish, K. (2024). *When Cohen's Kappa Is Not Enough: Exploring Methods for Estimating Inter-Rater Reliability for Time Sequential Data* [Manuscript submitted for publication].

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

Jansen, R. G., Wiertz, L. F., Meyer, E. S., & Noldus, L. P. J. J. (2003). Reliability analysis of observational data: Problems, solutions, and software implementation. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 391–399. https://doi.org/10.3758/BF03195516

Thomann, B. (2001). Observation and judgment in psychology: Assessing agreement among markings of behavioral events. *Behavior Research Methods, Instruments, & Computers*, *33*(3), 339–348. https://doi.org/10.3758/BF03195387

Walter, S. R., Dunsmuir, W. T. M., & Westbrook, J. I. (2019). Inter-observer agreement and reliability assessment for observational studies of clinical work. *Journal of Biomedical Informatics*, *100*, 103317. https://doi.org/10.1016/j.jbi.2019.103317