Observer agreement for timed-event sequential data: Time-based and event-based algorithms compared

Roger Bakeman[1], Vicenç Quera[2], and Augusto Gnisci[3]

[1]*Department of Psychology, Georgia State University, Atlanta, GA, USA, bakeman@gsu.edu*

[2]*Departamento de Metodología de las Ciencias del Comportamiento,*
Facultad de Psicología, Universidad de Barcelona, Barcelona, Spain, vquera@ub.edu

[3]*Dipartimento di Psicologia, Seconda Università degli studi di Napoli, Caserta, Italy, augusto.gnisci@unina2.it.*

Among observational researchers, a single data logging approach is becoming increasingly standard [1]. Working with digital multimedia recordings displayed on computer monitors, observers depress keys to note onsets of events. Offsets may be explicitly logged as well, or inferred from the onset of a later coded event in the same mutually exclusive and exhaustive (ME&E) set. With such instrumentation, continuously alert observers (continuous sampling) log data in a way that allows frequency, duration, co-occurrence, and contingency information to be derived later.

The present report uses computer simulation to compare five algorithms for assessing observer agreement given *timed-event sequential data* (TSD) [2], that is, continuously-sampled, time-logged observational data of the sort just described. Two are time-unit based: time-unit kappa and time-unit kappa with tolerance; and three are event based: The Observer algorithm, the INTERACT algorithm, and the Generalized Sequential Querier (GSEQ) dynamic programming (DP) algorithm, respectively. The first and second are implemented in GSEQ; the first and third in The Observer Version 5.0 [3], and the first in Mangold International's INTERACT. The fourth will be implemented in future versions of INTERACT and the fifth in future versions of GSEQ. The GSEQ DP algorithm is an extension of a dynamic programming algorithm we developed previously for *event sequent data* (ESD; only sequence but no times recorded) [4].

All algorithms are based on an agreement matrix (or confusion) matrix. The matrix is by itself useful for observer training because it shows how observers agree and disagree; and although all algorithms use the known formula to compute kappa, none satisfy the assumption of independent tallies required by the classic Cohen's kappa [5]. Thus the kappas produced should not be confused with Cohen's.

Algorithms

Time-unit based algorithms tally successive time units; if the time unit is a second, the kappa table contains 300 tallies for a 300 s observation. Time-unit kappa with tolerance ($\kappa_{\text{tolerance}}$) tallies an agreement if a match is found in the other observer's record, not just for the same second but within a stated tolerance (time-window, often of 2 time units). Because values vary slightly, depending on which observer is considered first, its value is computed as the mean of two values. Event-based algorithms link events and add tallies (agreements or disagreements) to the kappa table based on which events are linked.

Depending on the algorithm, some events may be linked to more than one other event, some may be linked to a nil event (one observer records a code the other does not, an omission-commission error), or some events remain unlinked.

The Observer algorithm is based on an algorithm described by Haccou and Meelis [6], the INTERACT algorithm is a modification of The Observer one, and the GSEQ DP algorithm is based on the classic Needleman and Wunsch (NW) algorithm [7] for aligning sequences of nucleotides, with modifications proposed by Mannila and Ronkainen [8] and additional modifications by us. The NW algorithm belongs to a broad class of methods known as *dynamic programming*, which permit exact solutions without exhaustively exploring myriad possibilities. Users specify costs for exact agreements, specific disagreements, and omission-commission errors; depending on these costs, the algorithm then determines an optimal alignment between two sequences, a backward trace through dynamic programming matrixes defined by the algorithm identifies agreement, disagreement, and omission-commission errors.

Simulation

We developed a simulation program (OASTES, or Observer Agreement for Simulated Timed Event Sequences) that generates master records and then simulates how observers might code those records. The program lets us vary the number of codes (*k*), the variability of their probability and duration, and the observer accuracy, and then computes kappa for the five algorithms. Kappas, averaged over 1000 simulations, were computed for k = 5, 10, and 15; for low, medium, and high variability; and for 75%, 85%, and 95% observer accuracy. Results are shown in Figure 1. Averaged over the circumstances simulated, $\kappa_{\text{tolerance}}$ tended to be higher and GSEQ DP kappas lower, with The Observer and Interact kappas intermediate. Kappa with tolerance, compared to without, averaged .06 higher.

Conclusion

Of the two the time-based algorithms, we prefer $\kappa_{\text{tolerance}}$, not necessarily because it gives higher values, as expected, but because we think it reasonable not to count minor errors of timing on the order of just a few seconds. Moreover, eliminating such errors from the agreement matrix leaves those disagreements which are arguably more serious, and which can profitably serve as a basis for further observer training.

Of the three event-based algorithms, we think the GSEQ dynamic programming algorithm is more accurate. The Observer and INTERACT algorithms do not allow for omission-commission errors, The Observer may link even quite distant events, and INTERACT leaves some events unlinked. We think they overestimate kappa, thus it is not surprising that they both produce higher values than the GSEQ algorithm for the circumstances simulated. Moreover, the Needleman-Wunsch algorithm, on which the GSEQ algorithm is based, is conceptually sophisticated and has a firm basis in the literature.

Time-unit based kappas, with a tally for each time unit, likely overestimates how often observers are making decisions, whereas event-based kappas, with a tally for each agreement, disagreement, omission, and commission likely underestimates the number of decisions observers make. Sometimes (perhaps often) observers decide that an event is continuing and not changing to another event; such agreements are not counted by the event-based algorithms—

indeed, how often these private mental events occur may be unknowable. We conclude with a simple recommendation, not either-or but both. Report values for both a time-unit kappa and an event-based kappa; this range likely captures the “true” value of kappa. Similarly, provide observers with agreement matrixes for both a time-unit and an event-based kappa. Each provides somewhat different (time-based vs. event-based) but valuable information as to how observers are disagreeing, and so are useful in different ways as observers strive to improve their agreement.

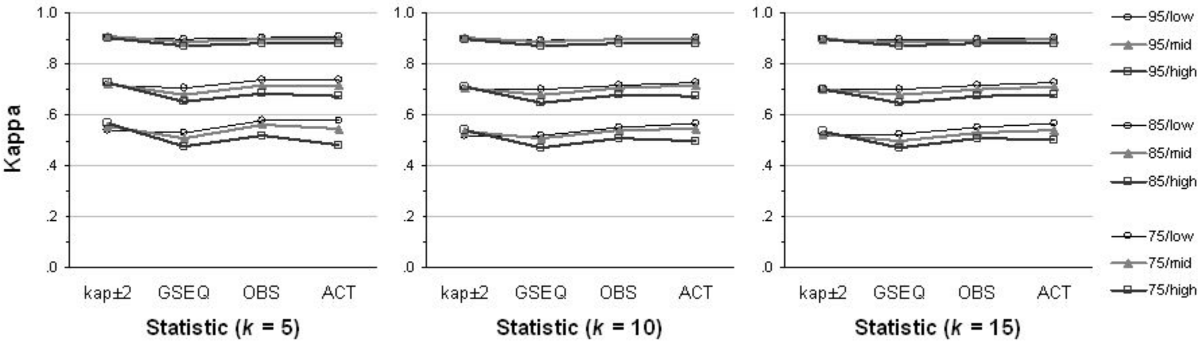

Figure 1. *Values for time-unit kappa (with 2 s tolerance), and as computed per the GSEQ dynamic programming, The Observer, and the INTERACT algorithms for k = 5, 10, and 15; observer accuracy = 75%, 85%, and 95%; and variability of code frequency and duration = low,* *moderate, and high.*

References

1. 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,** 391–299.
2. Bakeman, R., & Quera, V. (1995). *Analyzing Interaction: Sequential Analysis with SDIS and GSEQ*. New York: Cambridge University Press.
3. Noldus Information Technology (2003). *The Observer: Professional system for collection, analysis, presentation and management of observational data. Reference Manual, Version 5.0.* Wageningen, The Netherlands: Author.
4. Quera, V., Bakeman, R., & Gnisci, A. (2007). Observer agreement for event sequences: Methods and software for sequence alignment and reliability estimates. *Behavior Research Methods* **39,** 39–49.
5. Cohen, J. A. (1960). A coefficient of agreement for nominal scales. Educational *and Psychological Measurement* **20,** 37–46.
6. Haccou, P., & Meelis, E. (1992). *Statistical analysis of behavioural data: An approach based on time-structured models.* Oxford: Oxford University Press.
7. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48,** 443–453.
8. Mannila, H., & Ronkainen, P. (1997). Similarity of event sequences. In *Proceedings of the Fourth International Workshop on Temporal Representation and Reasoning. TIME’97* (Daytona Beach, Florida), 136-139

.