Journal of Data Science **5**(2007), 357-378

The Time Resolution in Lag-Sequential Analysis: A Choice with Consequences

André Berchtold[1] and Gene P. Sackett[2]
[1] *University of Lausanne and* [2] *University of Washington*

Abstract: The creation of data sets using observational methods for the lag-sequential study of behavior requires selection of a recording time unit. This is an important issue, because standard methods such as momentary sampling and partial-interval sampling, for instance, consistently underestimate the frequency of some behaviors. This leads to inaccurate estimation of both unconditional and conditional probabilities of the different behaviors, the basic descriptive and analytic tools of sequential analysis methodology. The purpose of this paper is to investigate the creation of data sets usable for the purpose of sequential analysis. We show that such data vary depending on the time resolution and that inaccurate choices lead to biased estimations of transition probabilities.

Key words: Behavior, bivariate distribution, data recoding, duration, lag-sequential analysis, time resolution, transition.

1. Introduction

In the process of recording behavioral data for purposes of unconditional or sequential analysis, an important step involves the choice of a common time reference for all observations. For instance, models such as Homogeneous Markov Chains and Double Chain Markov Models (Berchtold and Sackett, 2002) work better when each observation has the same duration. An observation can be defined as the data representing the behavior of a subject at a particular time, or the prominent activity during a 5- or 10-second period. The choice of the time resolution is generally based on issues that are seldom discussed in the methodology of a particular study. In this paper we show that the choice of the time reference can influence the results of a study, with different choices yielding different conclusions.

Many sampling methods have been used to collect behavioral observation data (Altmann, 1974; Suen and Ary, 1989). Prominent among these methods are Partial-Interval, Whole-Interval, and Momentary Sampling. In Partial-Interval

Sampling, the total session length is divided into a number of equal-length subintervals such as successive 15-second periods. Each behavior of interest is coded as one occurrence if it appeared at least once during the subinterval, regardless of the total number of actual occurrences, and zero otherwise. In Whole-Interval Sampling, the session also is divided into equal-length subintervals, but a behavior is coded as one occurrence only if it occurred continuously during the whole subinterval. In a variant of these two methods, each subinterval is divided into two units, an observing part and a recording part (Bijou, Peterson and Ault, 1968). For instance, a 15-second subinterval can be split into a 10-second observation period followed by a 5-second period during which the observer codes the behaviors that occurred during the preceding 10 seconds. In Momentary Sampling the subject is not observed continuously, but only at the end of a subinterval such as every 15 seconds, at which point a behavior is coded as having occurred if seen at this particular moment.

These methods and several variants have been widely used in the past (Kelly, 1977) and are still used in current research, as illustrated by the following examples. Kochanska, Coy and Murray (2001) used a modified whole interval method, measuring the predominant form of child compliance behavior occurring in 30-second segments of a test session. A major multi-site study of child care assessed 612 preschoolers using a momentary sampling method, measuring the presence or absence of peer social behaviors while observing for 30 seconds then recording for 30 seconds during 44 minute test sessions (NICHD Early Child Care Research Network). Partial interval 5-second time bins were used by Smith *et al.* (2002) to study conditional probabilities indexing responsiveness following reduced antipsychotic medication in people with intellectual disabilities. Robinson ét al. (2003) studied transitional behaviors leading to play by preschool children, using video tape to identify the single predominant social play activity occurring in successive 10-second periods, a modified whole-interval sampling procedure. Sexton, Hembre, and Kvarme (1996) used a 15-second interval to study Markov lagged probabilities within and between the behaviors of therapists and clients during psychotherapy sessions. The report did not provide sufficient detail to determine whether a partial or whole interval method was employed.

Several studies have shown that the use of sampling methods such as these can lead to biased results for the overall frequency and duration of observed behaviors and for the relation between successive behaviors. For instance, the frequency of rare behaviors is systematically underestimated by all kinds of time sampling methods, while their duration is overestimated by partial interval sampling and underestimated by whole interval sampling (Repp *et al.*, 1976; Harrop and Daniels, 1986; Suen and Ary, 1989). To correct this problem, the use of "real-time" recording methods has been advocated in which behavior is observed continuously and recorded in small time units such as tenths or whole seconds (Rapp *et al.*, 2001) . The result is a data file indicating when each behavior began and ended during a session. Computer and video technology has made this an easily implemented and reasonably accurate method for the production of behavioral data (Bakeman and Quera, 1995; Kahng and Iwata, 1998; Miltenberger, Rapp and Long, 1999; Thompson, Felce and Symons, 2000). However, as we show below, problems regarding time resolution can still occur.

Since it is possible to obtain data with a precision of one second or less, why is there any need to use longer time units or to collect data in long sampling bins in the first place? At least four reasons can be invoked.

1. The code may be too complex to score at a 1- or 2-second rate. Also, entry methods such as paper and pencil cannot be used for both closely watching and recording behavior at a fast rate. In either case, a long-time-unit method may be the only available coding method. Our results show that unless there are no behaviors with durations less than the sampling interval, such data are unlikely to provide good estimates of sequential probabilities (Sackett, 1978).

2. The time scored as the start or end of a behavior may not be the actual start or end time. Rather, this is the time when the observer noticed the event and then made the actual code entry. This raises the possibility for errors of several types. Error may occur if the observer needs time to realize that a behavior change has occurred, leading to late recording of the event. The degree of this error varies when it is difficult to distinguish one behavior from another, so the observer cannot tell exactly when a transition between behaviors took place. This error also varies because frequently occurring behaviors are often more easily detected than infrequent ones, even for experienced observers. This error will produce imprecise data when one uses brief sampling rates such as 1-second data. Recoding into longer time units could reduce the impact of this error source.

3. A short sampling rate is required when the study situation includes brief behavioral events. Also, a subject can perform an action for many seconds, then pause for a few seconds, then resume the same action for another long time. The problem then is to decide whether the pause should be recorded as a different behavior or as part of an ongoing action (maybe a “thinking” period). Recoding into a longer time unit could eliminate such brief pauses, but may also eliminate brief events of actual interest.

4. Many studies are designed so that the data can be compared with previously published data obtained with a long sampling rate. For such comparisons,

one must use either the same long, though potentially inaccurate, sampling rate, or recode a shorter rate.

In this paper we consider a data set in which the behavior of a focal subject was recorded in real time during socialization sessions with interactors. Focal behavior was coded into a set of mutually exclusive categories. Other information was also recorded, including the identification and behavior of interactors and the type of behavior as social or non-social. A new event was recorded each time either the behavior of the focal subject or the ID or behavior of an interactor changed. In raw form, each event can represent a different duration. A typical sequence of observations is shown in Table 1. This type of recording describes in detail the interaction between subjects, as well as a focal subject's non-social and self-directed behavior.

Table 1: Behavior of subject 1 at the beginning of the first socialization session as an example of the raw data. Each observation (row) consists of the behavior of the focal subject and of its interactor (0=Passive is entered as a dummy code for interactor on non-social and self-directed behaviors of the focal monkey), a numerical ID for each unique interactor or None for non-social actions and Self for self-directed activity, and duration of the event. These represent digits 2-4 of the original code. Digit 1, indicating whether a social behavior is initiated or received by the focal subject and if it is with or without physical contact, is not shown here.

Sequential observation	Focal behavior	Interactor behavior	Object-ID	Duration in seconds	Cumulative duration
1	Explore	Passive	None	4	4
2	Explore	Passive	Self	3	7
3	Explore	Passive	None	4	11
4	Explore	Passive	Self	3	14
5	Explore	Passive	2	5	19
6	Fear/Disturb	Play	2	8	27
7	Fear/Disturb	Explore	2	1	28
8	Passive	Passive	None	4	32
9	Fear/Disturb	Explore	2	6	38
10	Fear/Disturb	Play	2	4	42

A special difficulty for sequential analysis is that each sequential entry in Table 1 can have a different duration, so we begin by recoding the data into standardized 1-second events. Table 2 presents the resulting recoded file of the data shown in Table 1. Each data point now represents the behavior of the focal subject during exactly 1 second of the socialization session. This is the most decomposed and precise form of the raw data under the assumption that it takes an observer about one second to identify and record behavior during continuous "real time" observation.

Table 2: Recoding of the raw data of Table 1 into 1-second time units.

Second	Focal behavior	Interactor behavior	Object-ID	Second	Focal behavior	Interactor behavior	Object-ID
1	Explore	Passive	None	22	Fear/Disturb	Play	2
2	Explore	Passive	None	23	Fear/Disturb	Play	2
3	Explore	Passive	None	24	Fear/Disturb	Play	2
4	Explore	Passive	None	25	Fear/Disturb	Play	2
5	Explore	Passive	Self	26	Fear/Disturb	Play	2
6	Explore	Passive	Self	27	Fear/Disturb	Play	2
7	Explore	Passive	Self	28	Fear/Disturb	Explore	2
8	Explore	Passive	None	29	Passive	Passive	None
9	Explore	Passive	None	30	Passive	Passive	None
10	Explore	Passive	None	31	Passive	Passive	None
11	Explore	Passive	None	32	Passive	Passive	None
12	Explore	Passive	Self	33	Fear/Disturb	Explore	2
13	Explore	Passive	Self	34	Fear/Disturb	Explore	2
14	Explore	Passive	Self	35	Fear/Disturb	Explore	2
15	Explore	Passive	2	36	Fear/Disturb	Explore	2
16	Explore	Passive	2	37	Fear/Disturb	Explore	2
17	Explore	Passive	2	38	Fear/Disturb	Explore	2
18	Explore	Passive	2	39	Fear/Disturb	Play	2
19	Explore	Passive	2	40	Fear/Disturb	Play	2
20	Fear/Disturb	Play	2	41	Fear/Disturb	Play	2
21	Fear/Disturb	Play	2	42	Fear/Disturb	Play	2

It is possible to start an analysis using these data, but there are two potential problems. Even for well-trained observers, it is difficult to tell precisely when each behavior change occurs, and the use of a high time resolution such as 1 second can only increase this difficulty. This is a problem in computing observer reliability involving exact matching in real time (Bakeman *et al.*, 1997). Also, even with good rater reliability, the type of behavior can be in error. Such misclassification errors increase with increasing numbers of categories in the coding system. In response to problems such as these, we can analyze the data on a more aggregated temporal scale. The idea is to build a data set in which each observation has a fixed duration longer than that of the recoded 1-second data. Using this strategy, we can reduce the influence of short misrecordings, perhaps obtaining a more accurate picture of general trends in the data. Problems of assessing observer reliability also diminish, as observers have a longer time interval in which to display agreement.

However, recoding is not harmless. For instance, the total duration of rare events may be reduced even further. Moreover, identifying relationships between successive behaviors, the goal of sequential analysis, can also be artificially influenced. The problem is how to determine an optimal time base for the data.

Should each data point represent a period of 1, 5, 10 seconds, or some other duration?

This paper is linked to several important references in the literature. For instance, several papers focused on the accurate estimation of event durations with the use of post-hoc correction procedures (Suen and Ary, 1986; Quera, 1990). However, sequential analysis presents a very situation, since it requires the identification of two or more than two successive behaviors occurring in a particular order. So, recoding methods used for obtaining unconditional distributions may not be applicable to sequential analysis. In an important paper, Rogosa and Ghandour (1991) developed a powerful model for the analysis of behavioral data and related quantities such as overall frequencies and durations. They identified three main sources of error in collecting behavioral data: finite observations periods leading to undersampling of true distributions, observer errors, and heterogeneity over different observation periods. There is not much to add to the first two sources of error. All observed data sets are of finite length, so it is only possible to diminish the influence of this source of error by using the longest possible periods for collecting data, but it is not possible to completely suppress it. Observer errors are another important problem, the solution lying mainly in better training of observers. Rogosa & Ghandour also showed that increasing the number of simultaneous observers does not significantly improve the quality of the data. Finally, there is the problem of heterogeneity over different observation periods, or even within periods. Regarding this question, our approach is different from the one of Rogosa & Ghandour, due to our focus on dynamic models. While they see heterogeneity as a possible source of error which needs to be quantified through appropriate variance calculations, we consider heterogeneity to be a fundamental feature of behavioral data. The use of an appropriate statistical model can then both reveal the presence of heterogeneity and describe its dynamic. For instance, the Hidden Markov Model (Rabiner, 1989) is well suited for the analysis of unconditional distributions, while the Double Chain Markov Model (Berchtold and Sackett, 2002) focuses on dependence. The main message provided by both Rogosa & Ghandour and us is that many observational studies of behavior do not accord sufficient thinking to the method used to represent behavior sequences in time. This representation method is fundamental to obtain coherence between the goals of a study and the actual analyses, results, and conclusions.

In this paper we present a study comparing raw and recoded data by means of statistical tests to identify an optimal time base for analyzing continuous data. Our purpose is not to present a real analysis of this data set. Our intention is to simulate different sampling options and to compare them in regard to the behavior frequencies and transitions between behaviors which are reproduced. For the purpose of generalization, we also include a second set of comparisons based on simulated data. What follows is organized in three sections. First we present the characteristics of our raw data, the recoding method, and the test procedure used to compare different sampling intervals. Next we summarize our findings. We conclude with a discussion of the advantages and issues of different methods.

2. Methods

2.1 Subject characteristics and the observational coding system

We consider data from a group of 42 young pigtailed macaque monkeys (Macaca nemestrina), 21 males and 21 females. The subjects were nursery reared and experienced identical husbandry and caging methods (Ruppenthal and Sackett, 1992). They were separated from their mothers due to experimental requirements, premature delivery and/or low birth weight, injury or illness, maternal rejection, or illness or death of the mother. All infants had normal physical growth rates after month 2 and none were in any invasive prenatal or postnatal experiments.

Data collection followed a standard protocol and observational method (Ruppenthal and Sackett, 1992; Novak and Sackett, 1997; Worlein and Sackett, 1997). Data were collected during playroom socialization in groups of four or five infants between 16 and 363 days of age. The behavior of each subject as a focal individual and the behaviors of the interacting animals were observed for a randomly selected 5-minute period during daily 30-minute sessions. Data were recorded using a 4-digit observational code. Digit 1 coded the non-social or social nature of the current behavior, digits 2 and 3 coded nine mutually exclusive and exhaustive categories of behavior by the focal and interacting subject(s), respectively, and digit 4 coded the interactor ID for social behavior or various objects including self-directed actions for non-social behavior. A new entry was made for each change on any digit of the code. The method results in a record showing each sequential code and its duration in seconds, providing the raw data for analyzing overall frequencies, durations, and sequences of events. Observers were trained to between-observer reliabilities of kappa = .65 (Cohen, 1960) or better for agreement within ± 1 second on the total code.

In terms of frequency and duration, the four categories of Passive, Explore, Fear/Disturbance, and Play constitute over 98% of the behavioral repertoire of our nursery-raised infant monkeys (Worlein and Sackett, 1997). To simplify our presentation, we deal only with these four categories.

2.2 Recoding observations

Data in 1-second form, as presented in Table 2, are our most precise and detailed data, so they become the starting point for subsequent recoding. In an initial recoding, the behavior of focal monkeys was recoded into 5-second blocks. We considered two ways of doing that. The first one emulates a standard Momentary Sampling procedure. We assigned to each 5-second block the behavior observed during the first of the five seconds. The second procedure consists in replacing the five 1-second observations of each block by the behavior with the longest duration during the 5-second block. When two behaviors had the same duration, the first one occurring in that interval was coded for that block. Even if this method is not directly linked to the traditional Partial-Interval, Whole-Interval, and Momentary Sampling methods, it is interesting because it uses all available data. Moreover, the two methods are similar in studying agreement between several observers. For both recoding procedures, final session blocks that lasted fewer than 5 seconds were discarded. Table 3 presents the recoding of the Table 2 data using both the momentary sampling and the longest duration procedures.

Table 3: Recoding of the data of Table 2 into 5-second observations using two principles: Momentary sampling and longest duration.

Observation	Seconds	Focal Behavior	
		Momentary Sampling	Longest Duration
1	1 - 5	Explore	Explore
2	6 - 10	Explore	Explore
3	11 - 15	Explore	Explore
4	16 - 20	Fear/Disturb	Explore
5	21 - 25	Fear/Disturb	Fear/Disturb
6	26 - 30	Passive	Fear/Disturb
7	31 - 35	Fear/Disturb	Fear/Disturb
8	36 - 40	Fear/Disturb	Fear/Disturb

Table 3 shows that the two recoding procedures did not yield the same result. Momentary sampling seems better because the three different behaviors of the focal subject are still present in the recoded data, while only two are reproduced by the longest duration procedure. However, this finding is due to the small size of our example, so we cannot draw general conclusions yet. The point is that different sampling procedures can lead to different data distributions.

Following the same principle, other aggregations are possible using different block lengths. In this study, we considered four additional recodings, blocks of 2, 10, 15, and 20 seconds. Each recoding was performed starting from the raw 1-second data. Recoding the data into longer time units decreases the total number of observations and can lead to data sets that are too small for reliable analyses. In our study, the number of 1-second data points available for each subject ranged from 1844 to 25591 with a mean value equal to 14744 and a standard deviation equal to 4336. When the time intervals are of length $t > 1$ second, the number of data points is approximately the number of 1-second data points divided by t. In practice, the actual numbers of data points can be slightly less, because it is possible to loose some data at the end of each observational session when recoding into larger time units. The mean number of recoded data points actually used are 7362 (SD=2164), 2931 (861), 1460 (428), 970 (285), and 726 (213) for the 2-, 5-, 10-, 15-, and 20-second data, respectively.

2.3 Test procedure

Recoding into larger time units produces a loss of information. Therefore, it is necessary to determine whether the recoded data maintain the same characteristics as the original data. We compared observations of different lengths at two levels.

First, we determined whether recoding into longer time units influences the relative unconditional distribution of the four behaviors. To do that, we used the files containing 1-second data for each subject and the corresponding 2-, 5-, 10-, 15-, and 20-second data computed using the methods described above. After computing the distributions of the four behaviors for each data length, we compared each distribution with all other distributions corresponding to lower aggregation cases. For each subject, we first compared the 2-, 5-, 10-, 15-, and 20-second distributions with the 1-second distribution, then the 5-, 10-, 15-, and 20-second distributions with the 2-second one, and so on. The statistic used was a standard chi-square test at the 95% level. The lower aggregation distribution was used as a theoretical distribution, and the other was considered to be the observed distribution. The total number of data points was equal to that of the observed distribution. The null hypothesis specifies that both distributions are equal, its rejection signifying that the degree of aggregation of the observed distribution is too high to correctly reproduce the real distribution of the data. In other words, some frequencies estimated from the observed data are clearly different from the corresponding frequencies in the theoretical data.

For instance, the first subject's overall distribution for the 1-second data was

$$\begin{array}{ccccccc} & & \text{Passive} & \text{Explore} & \text{Fear/Disturb} & \text{Play} & \\ F_1 = & (& 6043 & 6422 & 797 & 2875 &) \end{array}$$

and for the 5-second data computed with the longest duration procedure

$$\begin{array}{ccccccc} & & \text{Passive} & \text{Explore} & \text{Fear/Disturb} & \text{Play} & \\ F_5 = & (& 1204 & 1295 & 156 & 551 &) \end{array}$$

Note that these distributions give the number of x-second data points associated with each of the four possible behaviors, not the actual frequency or duration distributions of the behaviors themselves. This is a characteristic of all of the time sampling methods discussed in this paper. Namely, these procedures do not yield true frequencies and durations, except under the restriction that one and only one behavior occurs in any interval (see Sackett (1978) for a more complete discussion of measurement units). However, 1-second data will represent a more precise approximation to real-time information than 5-second data. In our mutually exclusive and exhaustive observation system, one and only one behavior can occur per second, so both true frequency and duration units can be measured.

The expected number of data points corresponding to each behavior is obtained by multiplying each cell in the F_1 distribution by the total of the F_5 distribution, and dividing it by the total of F_1. For instance, we obtained the following expected distribution for the first subject:

$$\begin{array}{rccccccc} & & \text{Passive} & \text{Explore} & \text{Fear/Disturb} & \text{Play} & \\ T_1 = & (& 1200.59 & 1275.88 & 158.34 & 571.19 &) \end{array}$$

In this case, the hypothesis that the 5-second data distribution is statistically identical to the 1-second distribution was accepted (chi-square = 1.04 $< \chi^2(3, \alpha=.00122) = 15.85$). Notice that since we made 42 identical tests between each pair of data lengths, one for each of our subjects, we applied a Bonferroni correction to ensure the type I error to be globally equal to α=.05. Consequently, we made each individual test with a type I error fixed to α=.00122. The same procedure was used to compare 2-, 5-, 10-, 15-, and 20-second data with data of shorter time bases.

Second, we studied the influence of longer sampling intervals on the relations between successive behaviors. We began by computing the crosstables between every two successive data points for each subject (lag 1) and for each data length, and we applied the same χ^2 tests used in the case of the unconditional distribution of the four behaviors. Finally, for reasons discussed below, we performed two additional series of tests: χ^2 tests on the four diagonal elements of the crosstables, and χ^2 tests on the twelve non-diagonal elements of the same crosstables.

2.4 Theoretical experiment

To be complete and to check whether results obtained from our empirical data could be generalized, we performed another set of tests involving data generated by the mean of homogeneous first-order Markov chains. As before, we considered a random variable taking four different values. This variable follows a Markov chain $Q = [q_{ij}]$, $i, j = 1, ..., 4$, where q_{ij} is the probability of transition from state

i to state j defined as

$$q_{ij} = \begin{cases} \gamma\, U(0,1) & \text{if } i = j \\ U(0,1) & \text{otherwise} \end{cases}$$

where $U(0,1)$ is a randomly uniformly distributed variable on (0,1), and where γ ranges from 1 to 100. By increasing the probabilities located on the main diagonal of the matrix, the coefficient γ simulates different levels of autocontingency. For each value of γ, we computed 100 different transition matrices, and each matrix was used to generate a sequence of 1000 data points. We performed then the same chi-2 tests previously described in the case of the empirical data. Results are average computed on each set of 100 data sets corresponding to a value of γ.

3. Results

3.1 Unconditional distributions

Table 4 summarizes results concerning unconditional distributions, indicating the number of times the null hypothesis was retained, according to the χ^2 test, when comparing a given distribution with the distributions of shorter data lengths.

Table 4: Chi-square test results for the unconditional distribution of the four behaviors. The left part of the table concerns recoded data obtained with momentary sampling, and the right part of the table concerns data obtained with the longest duration procedure. The length of the reference observations is given in column and the length of the test observations is given in row. Cell numbers indicate how many of the 42 subjects had a good fit of the observed distribution to the shorter sampling rate reference expected distribution, as indicated by failure to reject the 3 degree of freedom χ^2 test. The type I error is globally set to .05 for each group of 42 tests with Bonferroni correction.

	Momentary sampling						**Longest duration**				
	1	**2**	**5**	**10**	**15**		**1**	**2**	**5**	**10**	**15**
2	42	-	-	-	-	**2**	42	-	-	-	-
5	42	42	-	-	-	**5**	41	41	-	-	-
10	42	42	42	-	-	**10**	41	41	42	-	-
15	42	42	42	42	-	**15**	37	37	41	42	-
20	42	42	42	42	42	**20**	34	34	41	42	42

We see from the table that the null hypothesis of similar distributions is always accepted in the case of momentary sampling, whatever the gap in time resolution between the reference and the test data. Results are slightly different for recoded data obtained with the longest duration method. With the 1-second data as the expected distribution, the 2-second recodings never differed and the 5- and 10-second recodings differed for only 1 subject. As expected, with an increasing time difference from the 1-second data, the probability of rejecting the null hypothesis increased. This occurred primarily because merging successive observations to produce longer intervals resulted in reducing more than proportionally the number of data points corresponding to short-duration behaviors. For instance, if Fear/Disturb is rarely observed for more than 5 seconds and we use a 20-second time resolution, it is likely that almost no observation will include Fear/Disturb. A consequence of this is, of course, that the relative number of data points corresponding to longer-duration behaviors will be overestimated. When considering the distributions of the 2-, 5-, 10-, and 15-second data as reference, the number of significant differences also increased with the difference between the reference and test time intervals. The shortest difference always provided better results, with only one χ^2 rejection for 5- against 2-second intervals, and no rejections for 10- against 5-second intervals, 15- against 10-second intervals, and 20- against 15-second intervals.

Results appearing in the right part of Table 4 are somewhat counterintuitive, because as we move down in the table the number of data points used for each χ^2 test decreases, reducing the power of the tests and so increasing the probability of accepting the null hypothesis. However, we observe that in fact the number of times the null hypothesis is accepted tends to decrease as we move to longer intervals. It appears that the decrease in power is more than counterbalanced by the increasing difference in longer observations between the reference and test distributions.

On the basis of these results, we conclude that recoding into somewhat longer observations can be done without much distortion of the “true” distributions, whatever the recoding method. However, using for instance 15- instead of 1-second intervals can produce unacceptable distortion with the longest duration method, significantly altering the distribution of 5 out of 42 subjects (11.9%). Note that many observational studies have collected data using 10- and 15-second intervals. To the extent that fairly short-duration behaviors were of interest, it is likely that the data of such studies distorted the actual distributions.

3.2 Crosstables

To analyze the influence of recoding upon sequential relationships, we constructed crosstables for the number of times a data point corresponding to a behavior was followed by every behavior including itself as the next data point (lag 1 data). The crosstable for the 1-second data of the first subject was

$$C_1 = \begin{array}{r|cccc|} & \text{Passive} & \text{Explore} & \text{Fear/Disturb} & \text{Play} \\ \hline \text{Passive} & 5170 & 534 & 40 & 277 \\ \text{Explore} & 469 & 5649 & 66 & 217 \\ \text{Fear/Disturb} & 67 & 43 & 644 & 38 \\ \text{Play} & 300 & 184 & 46 & 2340 \\ \hline \end{array}$$

where the first behavior occurring is in the row, and the subsequent behavior is in the column. For example, data points corresponding to the behavior Explore (row 2) were followed 217 times by data points corresponding to the behavior Play (column 4).

Table 5: Chi-square test results for the 16-cell sequential relationship crosstables between two successive behaviors. The left part of the table concerns recoded data obtained with momentary sampling, and the right part of the table concerns data obtained with the longest duration procedure. The length of the reference observations is given in column and the length of the test observations is given in row. Cell numbers indicate how many of the 42 subjects had a good fit of the observed distribution to the shorter sampling rate reference expected distribution, as indicated by failure to reject the 15 degree of freedom χ^2 test. The type I error is globally set to .05 for each group of 42 tests with Bonferroni correction.

	Momentary sampling				
	1	**2**	**5**	**10**	**15**
2	0	-	-	-	-
5	0	1	-	-	-
10	0	0	18	-	-
15	0	0	11	41	-
20	0	1	13	39	42

	Longest duration				
	1	**2**	**5**	**10**	**15**
2	0	-	-	-	-
5	0	1	-	-	-
10	0	1	29	-	-
15	0	1	28	42	-
20	0	3	31	42	42

Table 5 summarizes the results, indicating for how many of the 42 subjects the null hypothesis of the 15 degrees of freedom χ^2 test was retained. As before, the type I error was globally fixed to .05 for each set of 42 tests and a Bonferroni correction was applied. Expected frequencies for each of the 16 cells were calculated by multiplying the total frequency of the comparison table by the cell probabilities of the shorter length table. For both momentary sampling and longest duration methods the χ^2 test was always rejected when comparing every crosstable against the 1-second data. For the 2-second data, the test was also almost always rejected. The only substantial difference between the two recoding methods concerns the tests using the 5-second data as reference. In this case, tests obtained from momentary sampling were rejected about 67% of the time, while only 30% of the tests were rejected for data recoded with the longest duration method. Tests with the 10- and 15-second reference data led to almost no rejections for either method. When behaviors with durations shorter than 10 seconds are of interest, these results suggest that intervals as short as 5 seconds are unlikely to provide valid estimates of sequential behavioral relationships. The frequently used 10- or 15-second interval will markedly distort "true" sequential relationships for many subjects, even with as few as four categories constituting the behavioral repertoire under study.

The poor fit of sequential relationships from longer time intervals compared with short ones was expected for the following reason. Consider the crosstable C_1 above, computed on the 1-second data for subject 1 and the crosstable C_5 computed on the corresponding 5-second data obtained with the longest duration method:

$C_5 =$

	Passive	Explore	Fear/Disturb	Play
Passive	789	269	17	110
Explore	242	873	40	117
Fear/Disturb	35	29	74	15
Play	105	110	24	304

For a more easily understood comparison, we rescale C_1 into RC_1, which contains the same number of data points (3153) as C_5.

$RC_1 =$

	Passive	Explore	Fear/Disturb	Play
Passive	1013.5	104.7	7.8	54.3
Explore	91.9	1107.4	12.9	42.5
Fear/Disturb	13.1	8.4	126.2	7.4
Play	58.8	36.1	9.0	458.7

In the RC_1 data, the subject rarely switches every second from one behavior to another. Therefore, the frequency of staying in the same behavior (elements on the main diagonal of the crosstables) from one observation to the next is very high. When the data are aggregated as C_5, the number of transitions from one behavior to the same behavior decreases more quickly than the off-diagonal transitions. Thus, the 5-second off-diagonal data have amplified the relative frequencies identifying the switching process between different behaviors. A similar effect can be obtained by using the offset of events to determine the start of a new behavior, regardless of the duration of the preceding behavior (Sackett, 1979). This results in a proportionate decrease in the frequency with which a behavior follows itself, an undesirable outcome if autocontingency is of interest.

A simple example may better illustrate these recoding effects. Consider again crosstables C_1 and C_5 and suppose that we are interested in determining the behavior of a subject when he stops playing. For the 1-second data in row 4 of crosstable C_1, we see that the most common behavior following a Play data point (the conditional probability of behaviors following play) was Passive, which occurred $300/(300+184+46) = 56.60\%$ of the time, followed by Explore (34.7%) and Fear/Disturb (8.68%). With the 5-second data we see that the most common behavior following a Play data point was Explore (46.03%), then Passive (43.93%), and finally Fear/Disturb (10.04%). In this example, recoding not only modified the transition probabilities, but the sequential ordering was also transformed, leading to divergent conclusions.

Another way of analyzing the effect of recoding upon autocontingency is to compute the empirical equivalent of the autocontingency coefficient γ used in Section 2.4. A value of $\gamma = 1$ means that the average probability of transition from a behavior to itself is equal to the probability of transition from this same behavior to any other one. A value of $\gamma = 2$ means that the probability of transition from a behavior to itself is in average twice the probability of transition from this same behavior to any other behavior, and so on. Table 6 summarizes the results. Clearly, as the recoding becomes more extreme, the autocontingency coefficient decreases. Moreover, we observe that this phenomenon is much important with momentary sampling. So, the longest duration method should be chosen when autocontingency is the subject of interest.

Table 6: Empirical estimation of the autocontingency coefficient γ. The left part of the table concerns recoded data obtained with momentary sampling, and the right part of the table concerns data obtained with the longest duration procedure. The first column gives the length in seconds of the data. The other columns provide respectively the minimum value, the mean, the maximum value, and the standard deviation computed from the 42 subjects.

	Momentary sampling					**Longest duration**			
	Min	**Mean**	**Max**	**Std.dev**		**Min**	**Mean**	**Max**	**Std.dev**
1	4.61	19.35	49.97	10.49	**1**	4.61	19.35	49.97	10.49
2	2.78	8.96	23.55	4.87	**2**	2.78	8.96	23.55	4.87
5	1.93	4.59	23.17	3.45	**5**	2.21	5.13	26.42	3.79
10	1.48	3.20	15.54	2.25	**10**	1.75	4.00	18.19	2.59
15	1.14	2.78	11.01	1.70	**15**	1.65	3.98	14.47	2.23
20	1.12	2.56	9.60	1.46	**20**	1.61	3.77	12.42	2.03

We performed two additional sets of tests to illustrate the crosstable effects for all 42 subjects. First, we considered the four diagonal elements of the crosstables, that is the frequencies of staying in the same behavior from observation to observation. Table 7 summarizes the results of the corresponding 3 degree of

Table 7: Chi-square test results for the four diagonal elements of the crosstables between two successive behaviors, indicating the probability of remaining in the same behavior in successive intervals (autocontingency). The left part of the table concerns recoded data obtained with momentary sampling, and the right part of the table concerns data obtained with the longest duration procedure. The length of the reference observations is given in column and the length of the test observations is given in row. Cell numbers indicate how many of the 42 subjects had a good fit of the observed distribution to the shorter sampling rate reference expected distribution, as indicated by failure to reject the 3 degree of freedom χ^2 test. The type I error is globally set to .05 for each group of 42 tests with Bonferroni correction.

	Momentary sampling				
	1	**2**	**5**	**10**	**15**
2	37	-	-	-	-
5	16	27	-	-	-
10	14	23	42	-	-
15	22	27	42	42	-
20	20	27	42	42	41

	Longest duration				
	1	**2**	**5**	**10**	**15**
2	37	-	-	-	-
5	14	25	-	-	-
10	11	15	42	-	-
15	16	19	37	41	-
20	14	18	37	42	42

Table 8: Chi-square test results for the twelve off-diagonal elements of the crosstables between two successive behaviors, indicating the probability of switching behaviors between successive intervals. The left part of the table concerns recoded data obtained with momentary sampling, and the right part of the table concerns data obtained with the longest duration procedure. The length of the reference observations is given in column and the length of the test observations is given in row. Cell numbers indicate how many of the 42 subjects had a good fit of the observed distribution to the shorter sampling rate reference expected distribution, as indicated by failure to reject the 11 degree of freedom χ^2 test. The type I error is globally set to .05 for each group of 42 tests with Bonferroni correction.

	Momentary sampling				
	1	**2**	**5**	**10**	**15**
2	42	-	-	-	-
5	34	42	-	-	-
10	34	39	42	-	-
15	33	36	41	42	-
20	34	41	41	42	42

	Longest duration				
	1	**2**	**5**	**10**	**15**
2	42	-	-	-	-
5	30	40	-	-	-
10	32	36	42	-	-
15	34	35	40	42	-
20	32	35	40	42	42

freedom χ^2 tests. For both recoding methods, the null hypothesis was rejected in a large number of cases during comparisons with the 1- and 2-second data, and only comparisons against the 10- and 15-second data obtained good results for both methods.

Finally, we studied the 12 off-diagonal elements of the crosstables, performing 11 degree of freedom chi-square tests. The results, summarized in Table 8, are much better than those in Table 7, but the fit is poor for many subjects when $\geq$5-second blocks of data are compared with the 1-second data, and when $\geq$10-second blocks of data are compared with the 2-second data. We conclude that when considering sequential relationships, recoding data into longer time intervals produces extreme distortion on autocontingency, the probability of remaining in the same behavior. Recoding distortion is also present in off-diagonal elements, but the effect seems less severe.

3.3 Theoretical experiment

The results of the theoretical experiment are mostly similar to the results described above. In general, results tend to be better when the difference of aggregation between the theoretical and empirical data is small. Moreover, both aggregation methods (momentary sampling and longest duration) lead to very similar results. The comparison of unconditional distributions is always good, whatever the method of aggregation. We just note that momentary sampling performs slightly better when the data generating matrix has a low autocontingency coefficient γ.

On crosstables, results are poor when considering all cells, but the autocontingency coefficient γ plays an important role. Good results can be achieved when aggregation is moderated (2-seconds compared to 1-second, 5-seconds compared to 2-seconds, ...) and γ is very large (80 or above). On the other hand, results stay generally poor when γ is lower than 50, the typical values encountered in our behavior data (Table 6).

Except for data generated with very small values of γ (10 or less), results achieved on the diagonal elements of crosstables are very good. In the case of non-diagonal cells only, we have to consider values of γ larger or equal to 60 to achieve a majority of good results. These results on diagonal and non-diagonal cells seem to be in contradiction with the empirical results of Section 3.2, but there is a main difference between the two sets of data. In the Markovian generated data, the same value of the autocontingency coefficient γ is used for each row of the transition matrix, when this value can be very different for the real behavior data, ranging in one case from 9 to 168.7 for a same subject. Moreover, the empirical γ value is also very different from one subject to another as indicated by the minimum, maximum and standard deviation values of Table 6. Additional simulations showed that when the autocontingency coefficient is allowed larger variations between rows of the data generating matrices and between replications of the experiment, theoretical and empirical results become similar.

4. Discussion

In observational research, continuously recorded real-time data provide the most accurate and valid description of overall behavior durations as well as sequential relationships between behavioral events (Bakeman and Quera, 1995; Sackett, 1979). For practical purposes, "real time" can be operationally defined as the shortest time interval in which behavior can be coded reliably with a given methodology. When using a computer-assisted method to directly observe ongoing behavior, "real time" may be as short as 0.5-1 second. When observing from video recordings, a practical interval may be as short as 0.1 second, or even a single frame (0.033 second). With paper-and-pencil techniques, sampling intervals typically range from 5 to 20 seconds. The results presented here reveal problems of underestimation and overestimation of both behavior frequencies and transitions between successive behaviors when different sampling intervals are used to collect the data.

We considered two different methods for recording data, namely momentary sampling and longest duration. Although momentary sampling proved to work better for the estimation of unconditional distributions, neither of these methods appears to be better than the other for the comparison of crosstables. Momentary sampling seemed better when considering independently autocontingency and transitions between different behaviors (Tables 7 and 8), while longest duration worked better on complete crosstables (Table 5). In considering the two recoding methods, the major difference with regard to sequential analysis is that momentary sampling tends to break the relation between successive behaviours by taking into account only the behavior at a precise moment in time, while the longest duration method works in a smoother way, using all available information. So, even if momentary sampling works better when focusing on some subtype of events such as unconditional distributions, the longest duration method may provide a more general way of analysing transition processes.

Even if sampling in long time units can be justified, our results indicate that this methodology may produce more problems than solutions. The better data are always the most precise. The use of both an appropriate time unit and well-trained observers is the best answer to the issues discussed in the paper. One must determine a sampling interval short enough to accurately reflect the subject's behavior, yet long enough to yield a low rate of recording errors. This can be done by determining the duration of the shortest event of interest during a pilot stage of the research or from prior studies. Then this shortest duration can be used as an upper bound for the determination of the duration of each data point. For instance, if no event lasts for less than 5 seconds, then a 5-second sampling interval can approximate "real time".

It can be seen from the data in Tables 4-8 that the negative influence of longer time samples is roughly proportional to the difference between the time unit of the original and recoded data. Thus, the risk of distortion in the data increases with the size of the difference between the shortest "real time" interval and the interval used in a study. When a longer interval is necessary, this difference should be maintained as small as is practical. However, as seen in Table 5, even the difference between 1 and 2 seconds may be too large to tolerate the degree of distortion.

The behaviors we studied generally did not last for very long periods, so it is not surprising to observe that the overall transition process computed on 1-second data was significantly different from the one computed on 5- or 10-second data. Even when the same modelling technique is used to represent the transition process between data considered at different time resolutions, results can be very different. As an example, we fitted homogeneous Markov chains for each of our 42 subject and each time resolution. Using the Bayesian Information Criterion (Kass and Raftery, 1995), we determined that in most cases the first order chain was the best model for a subject, whatever the time resolution of the data. However, as can be deduced from crosstables C_1 and C_5, the transition processes computed from data with different time resolutions are actually very different, hence producing different transition matrices of the Markov chains. So, we cannot conclude that the analysis of the same data set at different time resolutions exhibits fractal properties, that is the same phenomenon being reproduced at different scales. On the contrary, each time resolution gives access to a different level of knowledge of the data and to different interpretations and conclusions. As showed by our results, one of the adverse effect of recoding is an important decrease of the autocontingency, this effect being higher with momentary sampling than with the longest duration procedure. So, recoding is clearly best suited for the analysis of behaviors lasting for long periods and it should be avoided when autocontingency is small.

The use of recoded data for the purpose of comparison raises several issues. We have seen that recoding can transform the meaning of the data. Also, even though results from recoded data may compare well with the results of other studies using the recoded sampling intervals, neither the recoded nor comparison data may accurately reflect the real behavior of the subjects. As a possible approach to this problem, consider a comparison study in which each data point represents exactly 10 seconds, and a new set of raw 1-second data. To compare results obtained from the new data set with the reference study, we could transform the

1-second data into 10-second data. However, even if we show that the recoded 10-second data yield the same results as those of the reference study, it would not mean that the more precise 1-second data describe the same phenomenon as the reference study. The solution is to perform two different comparisons: first compare the 1-second data with the recoded 10-second data, second compare the recoded data with the comparison study. If both comparisons indicate identical results, we can reasonably believe that conclusions from the new data are compatible with those of the comparison study. However, even in this case, we could not be certain what the conclusions would have been if the reference study had also used 1-second data.

For some purposes, we may be interested only in the relation between a subset of the behaviors or in a portion of their distribution such as the second half of each test session. However, distortions in the recoded distribution of some behaviors may be balanced by other recoded behaviors which were not distorted, leading to acceptable overall results of the tests. Thus, even if a particular recoding, or sampling rate of the primary data, does not seem to affect the data globally, it can have a distorting effect on some behaviors or at some but not all times during the observations. This means that testing for recoding distortion may need to be coordinated with the hypotheses or purposes of the study, necessitating a finer set of analyses than the global tests illustrated in this paper.

Acknowledgments

This work was supported by NIH grants RR00166 to the Washington National Primate Research Center and HD02274 to the University of Washington Center on Human Development and Disabilites. We thank Kate Elias for editorial assistance.

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour* **49**, 227-267.

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

Bakeman, R., Quera, V., McArthur, D., Robinson, B. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods* **2**, 357-370.

Berchtold, A. (2002). High-order extensions of the double Chain Markov model. *Stochastic Models* **18**, 193-227.

Berchtold, A., Sackett, G. (2002). Markovian models for the developmental study of social behavior. *American Journal of Primatology* **58**, 149-167.

Bijou, S. W., Peterson, R. F., Ault, M. H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis* **1**, 175-191.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37-46.

Harrop, A., Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis* **19**, 73-77.

Kahng, S. W., Iwata, B. A. (1998). Computerized systems for collecting real-time observational data. *Journal of Applied Behavior Analysis* **31**, 253-261.

Kass, R. E., Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association.* **90**, 773-795.

Kelly, M.B. (1977) A review of the observational data-collection and reliability procedures reported in the journal of applied behavior analysis. *Journal of Applied Behavior Analysis* **10**, 97-101.

Kochanska, G., Coy, K. C., Murray, K. T. (2001). The development of self-regulation in the first four years of life. *Child Development* **72**, 1091-1111.

Miltenberger, R. G., Rapp, J. T., Long, E. S. (1999). A low-tech method for conducting real-time recording. *Journal of Applied Behavior Analysis* **32**, 119-120.

Novak, M. F. S. X., Sackett, G. P. (1997). Pair-rearing infant monkeys (*macaca nemestrina*) using a "rotating-peer" strategy. *American Journal of Primatology* **41**, 141-149.

Quera, V. (1990). A generalized technique to estimate frequency and duration in time sampling. *Behavioral Assessment* **12**, 409-424.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257-286.

Rapp, J. T., Carr, J. E., Miltenberger, R. G., Dozier, C. L. and Kellum, K.K. (2001). Using real-time recording to enhance the analysis of within session functional analysis data. *Behavior Modification* **25**, 79-93.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F. and Berkler, M.S. (1976). A comparison of frequency interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis* **9**, 501-508.

Robinson, C. C., Anderson, G. T., Porter, C. L., Hart, C. H. and Wouden, M. (2003). Sequential transition patterns of social interaction during preschoolers free play: Is parallel-aware play a bi-directional bridge to other play states? *Early Childhood Research Quarterly* **18**, 3-21.

Rogosa, D., Ghandour, G. (1991). Statistical Models for behavioral Observations. *Journal of Educational Statistics* **16**, 157-252.

Ruppenthal, G. C., Sackett, G. P. (1992). *IPRL research protocol and technicians manual (2nd ed.).* University of Washington. http://www.rprc.washington.edu/iprl/contents.htm

Sackett, G. P. (1978). Measurement in observational research. In *Observing Behavior*, vol. 2 (Edited by G. P. Sackett), 25-34. University Park Press.

Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In *Handbook of infant development* (Edited by J.D. Osofsky), 623-649. New York:Wiley.

Sexton, H. C., Hembre, K., Kvarme, G. (1996). The interaction of the alliance and therapy microprocess: A sequential analysis. *Journal of Consulting and Clinical Psychology* **64**, 471-480.

Smith, C., Felce, S., Ahmed, Z., Fraser, W. I., Kerr, M., Kiernan, C., Emerson, E., Robertson, J., Allen, D., Baxter H., Thomas J. (2002). Sedation effects on responsiveness: evaluating the reduction of antipsychotic medication in people with intellectual disability using a conditional probability approach. *Journal of Intellectual Disability Research* **46**, 464-471.

Suen, H. K., Ary, D. (1986) A post-hoc correction procedure for systematic errors in time-sampling durations estimates. *Journal of Psychopathology and Behavioral Assessment* **8**, 31-38.

Suen, H. K., Ary, D. (1989). *Analyzing Quantitative Behavioral Observation Data.* Lawrence Erlbaum Associates.

Thompson, T., Felce, D., Symons, F. J. (ed.) (2000). *Behavioral Observation: Technology and Applications in Developmental Disabilities.* Brookes Publishing Co.

Worlein, J. M., Sackett, G. P. (1997). Social development in nursery-reared pigtailed macaques (*macaca nemestrina*). *American Journal of Primatology* **41**, 23-35.

Received February 24, 2006; accepted March 20, 2006.

André Berchtold
University of Lausanne
Institute of Applied Mathematics
SSP, Anthropole
CH-1015, LAUSANNE, Switzerland
Andre.Berchtold@unil.ch

Gene P. Sackett
University of Washington
Department of Psychology and National Primate Research Center
Box 357330, Seattle, WA 98195, USA
jsackett@bart.rprc.washington.edu