JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1980, 13, 493-500 NUMBER 3 (FALL 1980)

AN EMPIRICAL METHOD FOR DETERMINING AN APPROPRIATE INTERVAL LENGTH FOR RECORDING BEHAVIOR

R. W. SANSON-FISHER, A. DESMOND POOLE, AND JOHN DUNN

THE UNIVERSITY OF WESTERN AUSTRALIA

The study sought to examine the effects of varying interval length on the representation of data obtained using modified frequency time sampling. A 7-category scale was used to observe reliably the behavior of eight psychiatric inpatients. Using electronic real time recording equipment, it was possible to computer analyze the obtained data at varying interval lengths, the shortest interval being 1.0 seconds. It was found that increasing the interval length had little effect on the percentage of total duration recorded within each behavioral category, suggesting that this is a relatively stable measure of behavior. Percentage total events for each category was less stable with increasing interval lengths. The number of recorded events within each category tended to decrease, while their average durations tended to increase, as a function of increasing the interval length. The data suggest that the current practice of determining interval length in an arbitrary fashion, or on the basis of convention, should be abandoned. Rather, such a decision should be empirically determined for each particular observation scale and subject group. One method by which this might be achieved is presented.

DESCRIPTORS: interval length, within-interval error, empirical method of determination

Johnson and Bolstad (1973) and Jones, Reid, and Patterson (1975) have suggested that the development of direct observational techniques may well be the single most important contribution of applied behavior analysis to the discipline of psychology. Certainly such techniques are widely used by behavioral scientists.

Reviewing studies published in the *Journal of Applied Behavior Analysis* between 1968 and 1975, Kelly (1977) found that 76% employed direct observation procedures. Of the recording techniques used in those studies, 29% involved event recording, 20% interval recording, and 21% time sampling. Each of these observation tactics seeks to record data that accurately represent the actual stream of behavior being observed. However, even when the problems of interobserver reliability (Kazdin, 1977) are satisfactorily overcome, the results obtained using the different recording procedures are not necessarily comparable and all introduce some degree of distortion (Powell, Martindale, & Kulp, 1975; Powell & Rockinson, 1978; Repp, Roberts, Slack, Repp, & Berkler, 1976). In particular, Powell and Rockinson (1978) have demonstrated that interval recording procedures do not permit the frequency of discrete behaviors to be recorded accurately.

Interval recording, however, is widely employed as a means of representing a subject's behavior, but as pointed out by Jones et al. (1975) and Sanson-Fisher, Poole, Small, and Fleming (1979), this procedure creates serious problems for the researcher attempting to interpret the obtained data. This is primarily a result of the fact that in interval recording each interval is treated as representing a discrete behavioral event. Therefore, if the same behavior is coded in a sequence of intervals, it is impossible to

The authors extend their thanks to all staff and patients who permitted this study to be undertaken. The research was partially supported by a grant from The Sir Charles Gairdner Hospital Clinical Services Research Fund. Requests for reprints should be sent to the Secretary, Department of Psychiatry and Behavioural Science, University of Western Australia, Nedlands, Western Australia 6009.

determine whether this represents one continuous behavior or a number of sequential occurrences of that event. As Jones et al. (1975) state: "This difficulty of interpretation affects the scoring of the behavioral record, as, for example, in computing rates-per-minute, simple frequencies of occurrence, and probabilities of sequential interactions" (p. 55). Repp et al. (1976) have also demonstrated that with such procedures the utilized interval length can give rise to differences in the rates at which behaviors are reported to occur and in estimates of the duration of those behaviors.

The foregoing problems may be considered "across interval" errors, but modified frequency recording may also miss events if their duration is less than the employed interval length (Sackett, 1978). For example, consider the situation where an interval has been set at 6 sec and coding priority is determined by time dominance. If behavior A occurs for 4 sec, changes to behavior B for 3 sec, and reverts to the original behavior for a further 5 sec, behavior A will be recorded in both intervals. This loss of the short duration behavior B can be referred to as a "within-interval" error, which is a consequence of the interval length. The duration of the recording interval can, therefore, be of critical importance in determining the accuracy, or representativeness, of the obtained data.

Currently, the determination of the length of an interval appears to be largely an arbitrary matter rather than an empirical one. Jones et al. (1975) stated: "There do not seem to be any set rules about appropriate time samples. . . . The particular goals of the observational system probably define time sampling periods more appropriately than any procedural rules of thumb" (p. 54). As described by Jones et al. their group used a 6-sec modified frequency interval length, but Kelly (1977) reported that the most frequently used interval length, in studies reported in the *Journal of Applied Behavior Analysis,* is of 10-sec duration.

A search of recent behavioral studies indicated that no study specified the manner of selection, or appropriateness, of the utilized interval length even though it has been demonstrated that this has an impact on the accuracy of data (Powell et al., 1975; Repp et al., 1976). It is likely that an appropriate interval size, one which does not distort data, will be a complex function of the observation scale and the behavioral repertoire of the subjects (Repp et al., 1976). If this is correct, it suggests that prior to every experiment, and/or change of observation scale, some method of determining an appropriate interval length needs to be established.

A major advantage of empirically determining an appropriate interval length is that it may allow the calculation of discrete behavioral events by eliminating the occurrence of within-interval errors. In traditional interval recording, each interval has to be treated as containing a discrete event because it cannot be assumed that other behaviors did not occur for brief periods within an interval. Consequently, because of the possibility of such within-interval error, it is not necessarily justified to treat sequential intervals containing the same behavioral codes as representing one continuous event, although this may be what was actually observed. However, if within-interval error is eliminated, it is possible to sum across intervals to obtain a more accurate representation of discrete behaviors and their durations.

Systems that permit the recording of behaviors in "real time" eliminate within-interval error (Sanson-Fisher et al., 1979). Such systems are now possible due to recent developments in electronic and computer technology (Celhoffer, Boukydis, Minde, & Muir, 1977; Fitzpatrick, 1977; Sackett, Stephenson, & Ruppenthal, 1973; Sanson-Fisher et al., 1979; Stephenson & Roberts, 1977; Stephenson, Smith, & Roberts, 1975; Torgerson, 1977; Hollenbeck, Smythe, & Sackett, Note 1). According to Sidowski (1977), the main advantages of these systems are that they "allow the researcher to record the occurrences of behaviors as well as their durations in real time (allowing for subsequent serial and time-series analyses) and to produce outputs that al-

low for easy transfer of data to storage devices (e.g., magnetic tape or disk) or for direct entry into a computer" (p. 403). Another pragmatic reason for their use is the ease with which they can be employed by observers, who do not have to learn a pacing technique, as is necessary in interval recording procedures. Instead, observers need only to press a button to record a behavior's onset and termination (Sanson-Fisher et al., 1979).

However, even when data have been collected using "real time" recording procedures, analysis of the data still requires that an interval length must be set. The lower limits of such an interval is usually not imposed by the hardware or computer facilities. For example, the Data Acquisition in Real Time (DART 1) recording system described by Sanson-Fisher et al. (1979) allows for an interval length of .1 sec. At this interval length, data representation would appear to be more than adequate, given that it is unlikely that many behaviors of interest would occur for less than .1 sec. However, the need for human observers introduces constraints on the interval length that may be reliably used. This occurs as a result of such factors as the observers' need to glance at the equipment, their reaction times, the complexity of the rating scale, and the behavior under observation. Thus, although the use of very brief intervals is possible with "real time" hardware, the need to obtain observer reliability places constraints on the shortest interval that may be used.

The objective of the present study is to demonstrate a method to determine objectively an interval length at which little information of interest is lost, given a prescribed observational system and subject group. The study also provides an opportunity to examine the effect of differing interval lengths on the interpretation of collected data.

METHOD

Procedure

Video observations were carried out in an acute short-stay psychiatric unit described in detail by Sanson-Fisher, Poole, and Thompson (1979). Three video recorders were installed in the main patient areas over 3 months prior to beginning the study. During that period all patients admitted to the unit were informed that evaluation studies were taking place and requested to consent to being observed. Staff had consented to the research over 1 year previously. To further minimize observer reactivity, neither staff nor patients were aware of the nature of the observations being undertaken.

Throughout the study the video recorders were automatically switched on three times per day between 9:00 a.m. and 5:00 p.m. The eight target subjects were selected, at random, from all inpatients in the unit and were observed on eight separate 5 min occasions over 4 consecutive week-days.

Behaviors were coded by research assistants who were extensively trained as observers. The coding of behavior was carried out while observing the video recordings and using the DART 1 equipment (Sanson-Fisher et al., 1979). The inpatients' behavior was coded using the following 7-category observation scale.

Observation Scale

Positive self-concept. Behavior coded in this category reflects patients' positive self evaluation, optimism about achieving a satisfactory posttreatment adjustment and/or motivation for change, e.g., "I think I can handle my problems"; "I feel good/happy/content/relaxed."

Independent altruism. Asking questions about others' illnesses, their past experiences and future expectations, and offering solutions to another's problems were coded as independent altruism. Also included were comments that indicated patients were asserting their independence and requesting information about their condition, treatment, or other aspects of their psychiatric care, i.e., "How is your family?"; "Why have my drugs been changed?"

On task. All occasions on which patients were observed to be engaged in appropriate activities,

such as painting, pottery, woodwork, or reading were coded under this category.

Talk. General nonpsychiatric oriented conversation, not coded as positive or negative self-concept or independent altruism, was included in this category, e.g., talk about general issues, such as politics, movies, the weather, and other nonegocentric comments.

Negative self-concept. Comments reflecting a negative evaluation of self, life-style, coping skills, and derogatory conversations about close family members and friends were coded within this category, e.g., "I am worthless"; "I cannot cope."

Egocentricity. Behaviors included in this category were idle play and staring into space. Fixation on an object or person in a passive nonresponsive manner, self-stimulation, and sleeping were common examples of behavior coded as egocentric behavior.

Bizarre. Behavior coded bizarre was seen as inexplicable or irrational. It included such things as smiling, giggling, or weeping inappropriately; talking, muttering, or mumbling to oneself. Also included were bizarre movements, aggression in the absence of physical threat, claims of being controlled by other people or unusual forces, descriptions of phobic behavior and/or phobic avoidance behavior. The full definitions of this category were derived from traditional psychiatric scales.

In those situations in which the categories were not found to be mutually exclusive, previously determined priority rules were used. For example, if a patient was performing an on-task activity and also exhibiting bizarre behavior, the latter was coded.

The reliability of observations was assessed by having the same sequences of behavior independently coded by a second observer. Four 5-min sequences were randomly selected from observations made on days 1 to 3 and three from those obtained on day 4. The observers were not aware which sequences were selected for recording to assess reliability and, following the procedures recommended by Kazdin (1977), the sequences used for these checks were selected from the most complex available.

RESULTS

Reliability

Reliabilities were calculated separately for each observation category. Because of criticisms that percentage agreement, as a measure of reliability, fails to take account of chance agreements, the Kappa coefficient was used to estimate reliability (Hartmann, 1977). The levels of interobserver reliability, at the 1.0-sec interval length, are given in Table 1.

Data Analysis Procedure

The procedure for analyzing the data at the various interval lengths was identical throughout the study and will be described in relation to the 1.0-sec interval.

A specially developed software program scanned the data (stored on disk) on a second-by-second basis, and within each second determined which behavior dominated in terms of duration (i.e., time dominance). That is, if two behaviors occurred within a 1.0-sec period, behavior A for .4 sec and behavior B for .6 sec, behavior B was

Table 1

Level of interobserver agreement for each behavior category, at 1.0-sec interval length, using the Kappa coefficient.

Behavior category	*No. of reliability checks on which relevant category was coded*	*% occasions on which Kappa coefficient was significant at* $p < .05$
Positive self-concept	14	85.7
Independent altruism	8	87.5
On task	15	100.0
Talk	15	93.3
Negative self-concept	13	92.3
Egocentricity	12	83.3
Bizarre*	—	—

*Did not occur with sufficient frequency to calculate reliability.

coded as occurring within that interval. If behavior B again dominated within the next interval, it was treated as being the same event, and so on across all successive 1.0-sec intervals in which that behavior dominated. Once a different behavior dominated within a 1.0-sec interval, the occurrence of behavior B was considered to have terminated, and the duration of its occurrence was calculated, i.e., its duration equaled the sum of successive 1.0-sec intervals in which it dominated. Therefore, unless a behavior other than B occurred for less than .5 sec within an interval in this sequence, there was no within-interval error. Consequently, the representation of B as one event is an accurate representation of that behavior.

Using this procedure, the frequency with which each behavior category was recorded throughout the observation period was calculated, together with the average duration of those behaviors. The procedure was repeated using each of the longer interval lengths. However, as the interval increased so did the potential for within-interval error which is half the employed interval length, e.g., using the 5-sec interval behaviors occurring for less than 2.5 sec would be discounted in the determination of continuous behavioral events.

The Effects of Interval Length on the Representation of Behavior

Because 1.0 sec was the shortest interval at which satisfactory interobserver reliability could be achieved, these data were taken as criteria, i.e., the best estimate of the frequencies and durations of behaviors. The number of events and the average duration per event at 1.0 sec, together with those at interval lengths of 2.0, 3.0, 4.0, 5.0, 6.0, and 10.0 sec, are summarized in Table 2.

As can be seen, the effect of increasing the interval length is to decrease the number of occurrences of behavior recorded within each category. There is also a converse tendency for averaging duration per event to increase as a result of increasing the interval length. This effect is more clearly observed when the data obtained at each interval length are expressed as a percentage of those obtained at the criterion interval (i.e., 1.0 sec), as has been done in Table 3.

Table 2

Number of behavior events and their average duration for each behavior category as a function of interval length.

Behavior category	*Interval length*						
	1 sec	*2 sec*	*3 sec*	*4 sec*	*5 sec*	*6 sec*	*10 sec*
	NUMBER OF EVENTS						
Positive self-concept	55	49	39	34	31	27	18
Independent altruism	16	12	9	7	6	4	1
On task	219	193	171	155	143	134	108
Talk	80	66	59	54	44	38	28
Negative self-concept	53	50	49	43	42	40	35
Egocentricity	50	50	49	50	48	47	46
Bizarre	3	3	3	3	3	3	3
	AVERAGE DURATION PER EVENT (SEC)						
Positive self-concept	5.71	6.65	7.85	8.94	9.19	10.67	13.89
Independent altruism	2.94	3.17	4.00	5.14	6.67	6.00	10.00
On task	38.87	44.16	49.74	55.02	60.24	64.43	80.19
Talk	8.09	9.58	10.83	11.63	13.64	15.00	20.36
Negative self-concept	66.87	70.68	72.43	81.95	84.29	88.05	101.14
Egocentricity	108.02	108.20	110.39	108.48	112.50	115.28	119.78
Bizarre	144.00	143.33	145.00	142.67	143.33	144.00	146.67

Table 3

Number of behavior events recorded at each interval length as a percentage of those detected at 1.0 sec, and average durations as a percentage change from that obtained at 1.0 sec.

Behavior category	Interval length					
	2 sec	3 sec	4 sec	5 sec	6 sec	10 sec
	PERCENTAGE OF 1.0-SEC EVENTS					
Positive self-concept	89.1	70.9	61.8	56.4	49.1	32.7
Independent altruism	75.0	56.3	43.8	37.5	25.0	6.3
On task	88.1	78.1	70.8	65.3	61.2	49.3
Talk	82.5	73.8	67.5	55.0	47.5	35.0
Negative self-concept	94.3	92.5	81.1	79.2	75.5	66.0
Egocentricity	100.0	98.0	100.0	96.0	94.0	92.0
Bizarre	100.0	100.0	100.0	100.0	100.0	100.0
	PERCENTAGE CHANGE FROM 1.0-SEC DURATION DATA					
Positive self-concept	16.5	37.5	55.6	60.9	86.9	142.9
Independent altruism	7.8	36.1	74.8	126.8	104.1	240.1
On task	13.6	28.0	41.5	55.0	65.8	106.3
Talk	18.4	33.9	43.8	68.6	85.4	151.6
Negative self-concept	5.7	8.3	22.6	26.1	31.7	51.2
Egocentricity	.2	2.2	.4	4.1	6.7	10.9
Bizarre	−.5	.7	−.9	−.5	.0	1.9

This analysis indicates that the main effects of increasing the interval are to underestimate the number of events and to overestimate their average durations. It appears that, for the present data, the duration of an event has the greatest influence on the accuracy of its representation at the various interval lengths. Only in the case of the Bizarre category, which has an extremely long average duration per event, is there no loss of events and an extremely small percentage change in duration. Similar conclusions can be made about the Egocentricity category which is also characterized by a long average duration and again the effects of interval length appear to be small. On the other hand, Independent Altruism contains behaviors of short duration and, as a result, reflects the greatest variation with increasing interval lengths.

As interval recording techniques do not usually permit the reporting of a number of specific events, or their durations, because of the presence of within-interval errors, data are frequently reported in terms of either percentage total events or percentage total duration within each behavior category. The effects of varying the interval length on these measures are, therefore, summarized in Table 4.

As can be seen, the data obtained at the different interval lengths show little variation. However, Spearman rank order correlations were calculated between the ranks for percentage total events at the 1.0-sec interval and for those at the longer interval lengths, and a significant association ($p < .05$) between the 1.0-sec data and those at 2.0, 3.0, and 4.0 sec was found. This was not so for the other intervals. A similar analysis for percentage total duration revealed that the rank order correlations between the 1.0-sec data and those at the other interval length were throughout 1.0 ($p < .01$). It appears, therefore, that for the present data percentage total duration is the measure which is least susceptible to the effects of varying the interval length.

DISCUSSION

Consistent with previous research, the present study indicates that the choice of interval length may affect the accuracy with which observed behaviors are represented in the data. Although

Table 4

Percentage total events and percentage total duration, for each behavior category, as a function of interval length.

Behavior category	Interval length						
	1 sec	2 sec	3 sec	4 sec	5 sec	6 sec	10 sec
PERCENTAGE TOTAL EVENTS							
Positive self-concept	11.55	11.58	10.29	9.83	9.78	9.22	7.53
Independent altruism	3.36	2.84	2.37	2.02	1.89	1.37	.42
On task	46.01	45.63	45.12	44.08	45.11	45.73	45.19
Talk	16.81	15.68	15.57	15.61	13.88	12.97	11.72
Negative self-concept	11.13	11.82	12.93	12.43	13.25	13.65	14.64
Egocentricity	10.50	11.82	12.93	14.45	15.14	16.04	19.25
Bizarre	.63	.71	.79	.87	.93	1.02	1.26
r_s with 1.0 sec	—	.88*	.88*	.86*	.75	.68	.64
PERCENTAGE TOTAL DURATION							
Positive self-concept	1.66	1.73	1.62	1.61	1.51	1.52	1.32
Independent altruism	.25	.28	.19	.19	.21	.13	.05
On task	45.04	45.11	45.05	45.19	45.56	45.71	45.63
Talk	3.42	3.35	3.39	3.33	3.17	3.02	3.00
Negative self-concept	18.75	18.71	18.00	18.67	18.72	18.65	18.65
Egocentricity	28.58	28.64	28.65	28.74	28.56	28.68	29.03
Bizarre	2.29	2.28	2.30	2.27	2.27	2.29	2.32
r_s with 1.0 sec	—	1.00**	1.00**	1.00**	1.00**	1.00**	1.00**

*$p < .05$ (two-tailed).
**$p < .01$ (two-tailed).

the commonly used percentage events and percentage duration measures appear to be relatively unaffected by variations in interval length. this is not the case when information is required about discrete behaviors and their duration. This problem arises because of variability in the duration of discrete occurrences of the various categories of behavior. For example, in the present study, behaviors coded as Independent Altruism occurred for short durations (2.94 sec at the 1.0-sec interval length). Consequently, as the interval length was increased, an increasing number of such events were not recorded.

When the naturally occurring variability of different categories of behavior need to be accommodated within the constraints of an interval recording system, it appears necessary to determine the interval length so as not to lose short duration behaviors which may be of interest. By collecting data using a real time recording system, and then analyzing them at varying interval lengths, it is possible to examine the topography of the behavior so as to determine an optimum interval length, i.e., one at which within-interval error, caused by the missing of short duration behaviors, is minimized. This interval length may then be used to code behavior and still allow one to calculate discrete behaviors and their durations.

The results of the present study, for example, suggest that, given the same observation scale and subject group, the use of a 3.0-sec interval might be acceptable. Using this interval length, 79.6% of all events recorded at 1.0 sec are still represented in the data.

It is suggested that when data on frequency of events and/or their durations are required, the method described in this paper provides a possible technique for empirically determining an appropriate interval length.

REFERENCE NOTE

1. Hollenbeck, A. R., Smythe, L. E., & Sackett, G. P. BOSS: Behavioral Observation Scoring System—A

manual for computer-assisted observational research. Unpublished manuscript, University of Washington, 1975.

REFERENCES

Celhoffer, L., Boukydis, C., Minde, K., & Muir, E. The DCR-11 event recorder: a portable high-speed digital cassette system with direct computer access. *Behavior Research Methods and Instrumentation,* 1977, **9,** 442-446.

Fitzpatrick, L. J. Automated data collection for observed events. *Behavior Research Methods and Instrumentation,* 1977, **9,** 447-451.

Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis,* 1977, **10,** 103-116.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts and practice.* Champaign, Ill.: Research Press, 1973.

Jones, R. R., Reid, J. B., & Patterson, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 3). San Francisco: Jossey-Bass Inc., 1975.

Kazdin, A. Artifact, bias, and complexity of assessment: The ABC's of reliability. *Journal of Applied Behavior Analysis,* 1977, **10,** 141-150.

Kelly, M. B. A review of the observational data-collection and reliability procedures reported in the *Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis,* 1977, **10,** 99-101.

Powell, J., Martindale, A., & Kulp, S. An evaluation of time-sampling measures of behavior. *Journal of Applied Behavior Analysis,* 1975, **8,** 463-469.

Powell, J., & Rockinson, R. On the inability of interval time sampling to reflect frequency of occurrence data. *Journal of Applied Behavior Analysis,* 1978, **11,** 531-532.

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

Sackett, G. P. Measurement in observational research. In G. P. Sackett (Ed.), *Observing behavior: Data collection and analysis methods* (Vol. 2). Baltimore: University Park Press, 1978.

Sackett, G. P., Stephenson, E., & Ruppenthal, G. G. Digital acquisition systems for observing behavior in laboratory and field settings. *Behavior Research Methods and Instrumentation,* 1973, **5,** 344-348.

Sanson-Fisher, R. W., Poole, A. D., Small, G. A., & Fleming, I. Data acquisition in real time—an improved system for naturalistic observation. *Behavior Therapy,* 1979, **10,** 543-554.

Sanson-Fisher, R. W., Poole, A. D., & Thompson, V. Behaviour patterns within a general hospital psychiatric unit: An observational study. *Behaviour Research and Therapy,* 1979, **17,** 317-332.

Sidowski, J. B. Observation research: Some instrumental systems for scoring and storing behavioral data. *Behavior Research Methods and Instrumentation,* 1977, **9,** 403-404.

Stephenson, G. R., & Roberts, T. W. The SSR system: A general encoding system with computerized transcription. *Behavior Research Methods and Instrumentation,* 1977, **9,** 434-441.

Stephenson, G. R., Smith, D. P. B., & Roberts, T. W. The SSR system: an open format event recording system with computerized transcription. *Behavior Research Methods and Instrumentation,* 1975, **7,** 497-515.

Torgerson, L. Datamyte 900. *Behavior Research Methods and Instrumentation,* 1977, **9,** 405-406.

Received March 16, 1979
Final acceptance February 22, 1980