12-1975

An a Priori Analysis of the Internal Validity of Interval Recording

Christopher R. Milar
Western Michigan University

AN A PRIORI ANALYSIS OF THE INTERNAL VALIDITY OF INTERVAL RECORDING

by

Christopher R. Milar

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment
of the
Degree of Master of Arts

Western Michigan University
Kalamazoo, Michigan
December 1975

ACKNOWLEDGMENTS

I would like to thank Dr. Robert P. Hawkins for his intermittent reinforcement throughout the writing of this thesis. To Dr. Howard Farris and Dr. Jack Michael, I extend my sincere gratitude for their constructive criticism and encouragement during my oral examination. I would also like to thank Katharine Milar for her support during the numerous rewritings of this thesis.

Christopher R. Milar

TABLE OF CONTENTS

INDEX OF FIGURES

INTRODUCTION

The collection of data in quasi-natural and natural settings has become an integral part of most applied studies in behavior analysis. Due to the nature and purpose of these studies--modification of complex human behavior in the natural environment--it has often been difficult, if not impossible, to maintain experimental control of the degree found in the laboratory (Baer, Wolf, and Risley, 1968). Because of the interest in socially significant responses, recording of behavior is not typically achieved by electromechanical relay circuits and the like, but rather by human observers using a verbal (as opposed to electromechanical) definition of the behavior.

One of the most popular observational recording methods is interval recording. A review of the articles pertaining to classroom settings published in the Journal of Applied Behavior Analysis (Volumes I through V) shows that interval recording has been used in approximately 40% of these studies.

The typical procedure when using any observational recording method is to first define the behavior to be recorded in terms of observable events, then train observers in the use of the definition and the recording technique, and finally record data on the behavior of interest. With interval recording the observers commonly use a recording sheet similar to the one illustrated in Figure 1. The data sheet is marked off in small squares representing time inter-

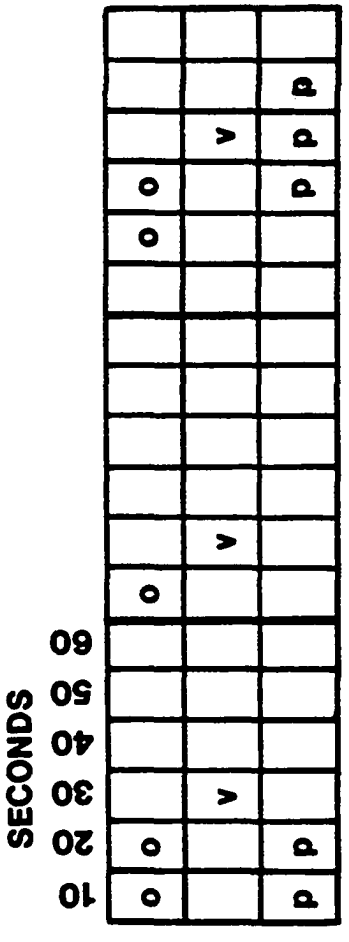
vals of equal size, often 10 seconds (see Bijou, Peterson, and Ault, 1968). The observer then records the occurrence or non-occurrence of the behavior for each successive interval of the session using any convenient symbol to represent the occurrence of the behavior. A stop watch is typically used to determine the passing of intervals but audible-signal generating devices have also been used (Worthy, 1968). Usually the behavior is recorded as occurring if it is taking place during any portion of the interval, no matter how small that portion.

Bijou, Peterson, Harris, Allen, and Johnston (1969) and Mattos (1971) have pointed out certain precautions regarding interval recording, and Hawkins and Dotson (1972) have demonstrated that there are serious shortcomings in the most popular method of calculating inter-observer agreement (reliability) scores from interval data. However, it appears that behavior analysts have not systematically examined the effects of using the interval recording method on the validity of the data obtained.

In a sense, there are two kinds of validity requiring examination when interval recording is used for data collection. First is the question as to whether the absolute level of a behavior is represented accurately by this method. For example, when one uses the frequency of a response as the criterion for assessing validity, interval recording is very likely to be at least somewhat invalid, in the absolute sense, because either of two errors can occur: there may be two or more occurrences of the response within a single

Figure 1

A sample section of a data sheet used with interval recording. Each column of 3 cells represents one 10-second interval. The letters in each row indicate that the particular response was occurring during some portion of that interval.



interval, and the recorded data will show only one event; or a single response may endure for so long a duration as to cover several intervals, and the data will show several events where there was only one. But where such absolute-level invalidity is present there may also be serious threats to another, perhaps more important type of validity; that which is referred to by Campbell and Stanley (1966) as the internal validity of the experimental effect. In examining internal validity one is raising the question whether the independent variables employed truly had the effect represented by the data. Or, as Campbell and Stanley put it, "Did in fact the experimental treatments make a difference in this specific experimental instance (p. 5)?" Obvicusly, if we cannot believe that the effects we see in our data are real, scientific experimentation is, at best, wasted. To the extent that the effects are misrepresented in magnitude, and particularly, where they are exaggerated, our enthusiasm about our results must at least be cautious.

Campbell and Stanley (1966) indicate several variables which could interact with the dependent variable so as to threaten internal validity. One of these is referred to as "instrumentation" and includes such phenomena as changes in the method of measuring the dependent variable during the experiment, changes in the calibration of the measuring instrument, and changes in the observers' recording of the behavior. If interval recording does affect the internal validity of results, it would appear to be a problem of

"instrumentation."

The purpose of the present analysis is to assess the threat to both of the above kinds of validity accruing from the fact that an experimenter elects to employ interval recording. At least four variables (which are not mutually exclusive) can be identified that are always present in experimentation and which could affect the accuracy of the data recorded by the interval method. The first two of these variables are the rate and the duration of the behavior. Examples of possible validity problems associated with high frequency or long duration responses were given above. Although frequency and/or duration are the primary data for most experimental studies, their relationship to the accuracy of the data recorded by the interval method has received little attention, with the exception that Mattos (1971) has stated that interval recording is generally inappropriate for behaviors with very high or very low rates.

The third variable which might affect the accuracy of interval-recorded data is the size of the interval used to record the behavior in question. The effect of the size of the interval on the accuracy of interval recording was raised as early as 1939 (Arrington, 1939). Using the interval method (then called time sampling) with 5-second intervals, Arrington collected data on the "speech episodes" of kindergarten girls. She ranked the girls as to the total frequency of speech episodes and then retabulated the data as though they had been recorded with 10-, 15-, 30-, and

60-second intervals. Her results showed that for the subjects with the highest ranks (4.25 speech episodes/minute) and the lowest (0.46 speech episodes/minute) there was little change in rank order with change to different size intervals. However, the subjects ranking around the median showed considerable change of rank with changes in the length of the recording interval. Her conclusion regarding intervals above 15 seconds in size was that:

> As the size of the interval increases further . . . the measures of total frequency become less discriminative and the ranks deviate more widely from the original five second ranking (p. 171).

However, it should be noted that her results could have been obtained by a simple mathematical analysis, without measuring real behavior, and that this approach would permit a complete parametric study of the relationship between interval size and the validity of the data at any selected response rate.

Bijou et al. (1969) also recognized the effect of interval size on the data obtained. They recommended the use of small intervals when recording a high rate behavior. Also, Mattos (1971) recommended that the length of the interval should be short enough to prevent multiple, unrecorded occurrences of the behavior within an interval.

The fourth variable which could affect the accuracy of interval recording is the temporal patterning of the behavior, which can be expressed in terms of the magnitude and variability of inter-response times (IRT). This variable appears to have been completely neglected in the literature, although Mattos (1971) does allude to

some interaction between IRT and interval size.

The present research, then, was an investigation of the effect of four variables--rate, duration, size of the recording interval, and IRT--on the accuracy of data collected by the interval method. The research employed a mathematical analysis and was oriented toward assessing threats to the internal validity of experimental data.

AN ANALYSIS OF THE INTERNAL VALIDITY OF DATA OBTAINED BY THE INTERVAL RECORDING METHOD

In any scientific study of a natural phenomenon, the investigator attempts to isolate the effects of various independent variables and to assess the relationships which exist between the independent and dependent variables. In the present analysis the independent variables are three response dimensions--rate, duration, and IRT--and one measurement dimension--the size of the recording interval. The dependent variable was the accuracy of the data obtained by interval recording. Accuracy was defined as the degree to which obtained results matched those that would have been obtained by a frequency and/or duration measurement of the same behavior.

Due to the nature of the independent variables, attempts to fully assess their effects on the validity of the recorded data would not be feasible in the natural environment. For example, if one were collecting data on the behaviors of a child in a classroom, it would be difficult, if not impossible, to manipulate the frequency, duration, and IRTs of the behavior. In addition, any study conducted in such a setting would require observers to record the data, and when observers are involved, there is usually no true measure of the behavior to use as a validity criterion. One expedient way of evaluating the accuracy of interval recording is to carry out a mathematical analysis of the effects of the independent

variables on the accuracy of hypothetical data. This procedure also allows for a parametric analysis that would be impractical in a real life setting.

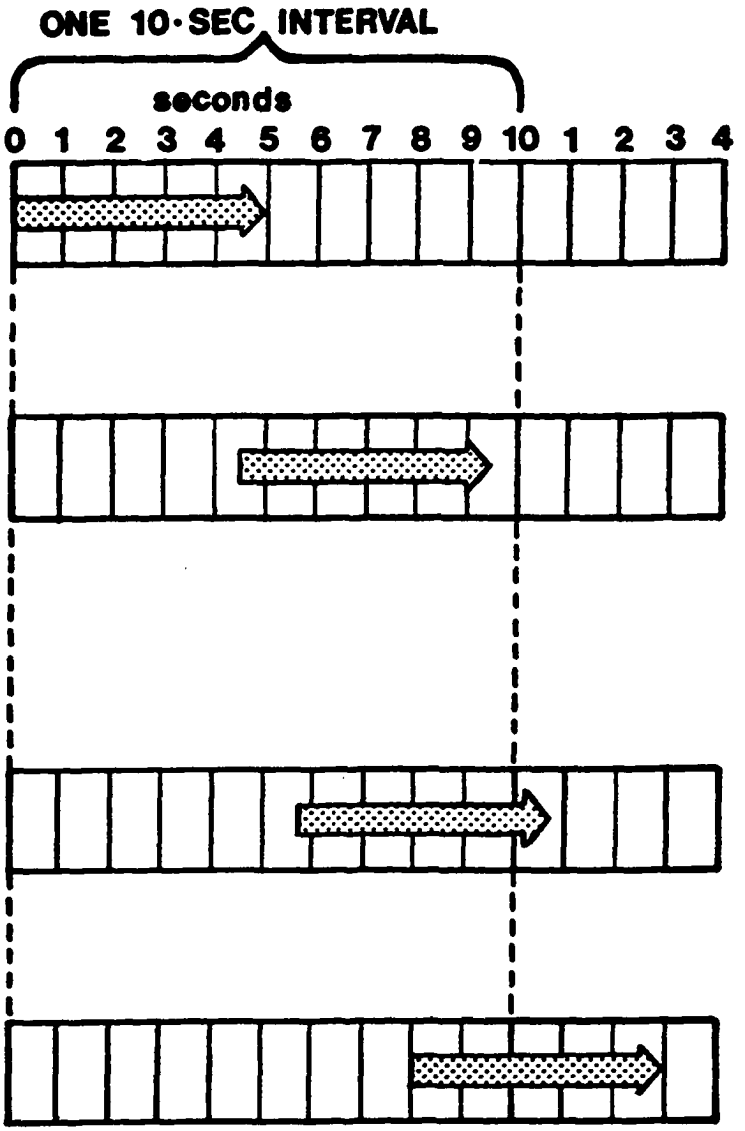
Method

The initial step in the analysis is to derive some simple probabilities. To begin the derivation, if a behavior has a duration of 5 seconds and occurs once during a 30-minute recording session, according to probability theory, it is equally likely that the behavior will commence at any second of the 10-second interval. If it commences at any time between the initiation of the interval (0 seconds) and the midpoint of the interval (5 seconds) the behavior will be recorded as occurring in only that interval. However, if the behavior commences after the 5-second point, it will be recorded as occurring in 2 intervals. Figure 2 illustrates two instances of each of these events, with the arrows representing a behavior which is 5 seconds in duration. Thus, there is a probability of 0.50 that a 5-second behavior will occur at such a point in time that it will be recorded in one interval and a probability of 0.50 that it will occur so as to be recorded in two intervals.

Likewise, if a behavior has a duration of 2 seconds and occurs once during a 30-minute recording session, the probability that it will be recorded in only one interval approaches 0.80 and the probability that it will be recorded as occurring in 2 intervals approaches 0.20. A behavior with a duration of 10 seconds which

Figure 2

Occurrences of a 5-second behavior within a 10-second recording interval. The arrow represents the behavior.



occurs once during a 30-minute recording session will, using 10-second intervals, have a probability approaching 1.00 of being recorded in 2 intervals, and a probability approaching 0.00 of being recorded in one interval.

Once the probability of each possible outcome (e.g., the behavior's being recorded in one interval or two) from a single occurrence of the response is known, it is possible to determine the most probable outcome from any selected number of occurrences of the response. For example, if a 5-second response has equal probability of being recorded in one or two intervals (both 0.50), on repeated occurrences the response will tend to occur in two intervals the same number of times as it occurs in one interval (provided the time between the responses, or IRT, is not less than the length of the recording interval). Then with two occurrences of the response the most probable outcome is that it will be recorded in a total of 3 intervals; with four occurrences, the most probable outcome is that it will be recorded in a total of 6 intervals; with twenty occurrences, the most probable outcome is that it will be recorded in 30 intervals.

Similarly, with a 2-second response occurring, say, ten times, the most probable outcome is that on two of those occurrences it will be recorded in two intervals and on the other eight occurrences it will be recorded in one interval; thus, the ten responses will be recorded in 12 intervals. With fifteen occurrences the most probable outcome would be the behavior's being recorded in 18

intervals.

The mathematical operations being employed here are identical to those presented by Hays and Winkler (1970) under the concept of mathematical expectation. Mathematical expectation is simply a weighted arithmetic mean based on the probabilities of the various possible outcomes. Hays and Winkler explain expectation as follows:

> . . . this term has been retained in mathematical statistics to mean the long-run average value for any random variable over an indefinite number of samplings. . . . Over a long series of trials, we can "expect" to observe the expected value (p. 137).

The formula these authors provide is

$$aE(X) = a\ xP(x)$$

$$= a\ x_1P(x_1) + x_2P(x_2)$$

where E(X) is the number of intervals in which the behavior is expected to occur, a is the frequency of the behavior, and x_1 and x_2 are the two possible numbers of intervals in which the behavior could be recorded. Given a particular interval size (say 10 seconds), x_1 and x_2 depend completely on the response duration. If the behavior has a duration of 5 seconds, x_1 is the occurrence of the behavior in only one interval and x_2 is the occurrence of the behavior in two intervals. If the behavior has a duration of 20 seconds, x_1 is the occurrence of the behavior in two intervals and x_2 is the occurrence of the behavior in three intervals. P is the previously calculated probability of these events.

Applying this formula to multiple occurrences of an event, suppose that an observer were recording a behavior with a duration

of 5 seconds and the behavior occurred four times during a 30-minute recording session. In this case, the expected value would be:

$$aE(X) = 4\ 1(0.5) + 2(0.5)$$
$$= 4\ 1.5$$
$$= 6$$

That is, over a series of sessions this behavior would be expected to be recorded in an average of 6 of the 10-second intervals per session. If the same 5-second behavior occurred with a frequency of 20 during a 30-minute recording session, the behavior would be expected to be recorded in 30 intervals. Of course the expected value does not represent the only value that can result. It represents the mode of a distribution of possible outcomes and only in the long run will this value be approximated.

Now that the foundation has been laid, the analysis can be applied to behavior changes as they would be reported in the literature. For the purpose of this section of the analysis, the interval size and the IRT will be held constant. The interval size will be 10 seconds and the IRTs will always be equal to or greater than the length of the interval used to record the behavior (10-second IRT for 10-second intervals), thus preventing two responses from occurring in the same interval.

The effect of frequency and duration

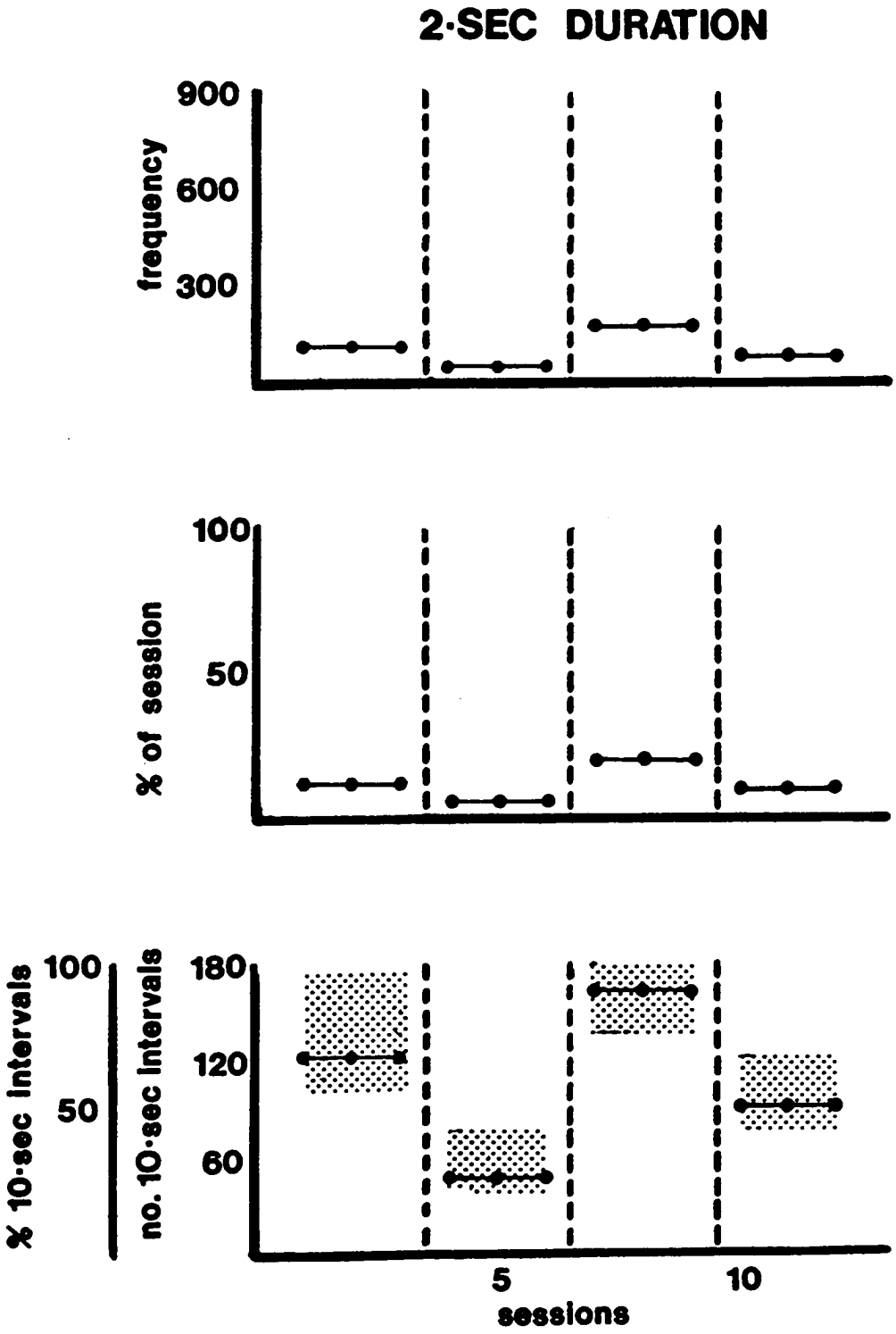
Figure 3 is designed to resemble the results of an experiment

employing the ABAB design often found in behavior analysis. The actual data are hypothetical, but all three graphs represent the same data obtained on a 2-second duration response during a 30-minute session. The top graph presents the frequency of the response under each of the four phases of the experiment. The middle graph depicts the total duration of the same behavior represented as the percent of the session in which the behavior was occurring. The bottom graph will be discussed shortly. The first three data points in the top graph show the behavior occurring with a frequency of 100 per 30-minute session. If the duration had been recorded, it would have been found that the response consumed 11% of each session, as shown in the middle graph. Thus, if the ordinate of these two graphs is extended to the maximum possible value (900 responses or 100% of the session) the results will appear the same on both graphs.

But what results would the experimenter obtain if he had used interval recording? In the first phase of the experiment the behavior could have been recorded in as few as 100 intervals (56% of the intervals) or as many as 180 intervals (100% of the intervals), depending on where the response happened to fall in relation to the interval boundaries (as described earlier and depicted in Figure 2). This range of possible results is represented by the shaded area in the first section of the bottom graph. If the expected value formula is applied we find that the most probable outcome is that the behavior would be recorded in 120 intervals (67%), as depicted by

Figure 3

Hypothetical results from an ABAB experimental design yielding four different frequencies of a behavior with a duration of 2 seconds. The top section shows the actual frequency of the behavior. The middle section shows the percent of session in which the behavior occurs. The bottom section shows the percent of intervals (or number of intervals in a 30-minute session) in which the behavior is expected to be recorded. The shaded area is the range of possible values.



by the data points in the first section of the bottom graph.

Suppose the frequency of this response were then changed from 100 to 38 per session as is shown in the second section of the top graph of Figure 3. The middle graph shows that the response would then occur during only 4% of the session. However, as the second section of the bottom graph shows, interval recording would result in the response's being recorded in at least 38 intervals (21%) or as many as 76 intervals (42%), with the most probable being 45 intervals (25%).

The third section of the top graph in Figure 3 shows an increase in the frequency of the behavior to 135 per 30-minute session. The middle graph shows that the behavior now occurs during 15% of the session. The interval method results in the behavior's being recorded in as many as 180 intervals (100%) and as few as 135 intervals (75%), with the most probable outcome being 162 intervals (90%).

The fourth section of the two top graphs in Figure 3 shows that if the response were then reduced in frequency to 75 per session, it would consume 8% of the total session. In this case the response would be recorded by the interval method in at least 75 intervals (42%) and as many as 150 intervals (84%). The most probable number of intervals would be 90 (50%).

Thus, the analysis appears to indicate that not only does interval recording distort the absolute level of the behavior (as it must when IRT equals or exceeds the interval size, due to the

"crediting" of a full interval any time the response occupies some fraction of the interval), but it can also distort the magnitude of the behavior change. In Figure 3 the actual behavior change between condition 1 and condition 2 is 62 responses or 124 seconds of the behavior, each of which constitutes a change of only 7% of the maximum possible frequency or duration. Yet, the interval data could show a change as large as 79% of the session (142 intervals) or as small as 14% of the session (24 intervals). The same exaggeration of experimental effect is seen in every behavior change represented in Figure 3.

Figure 4 shows hypothetical data from an experiment involving a response of 5-second duration. As with the 2-second response (Figure 3), interval recording exaggerates both the absolute level of the response and the magnitude of the behavior change reported. Figures 5 and 6 show similar hypothetical experiments with responses of 15-second and 30-second duration respectively.[1]

It should be noted that the exaggeration of both the absolute level of the behavior and the distortion of the experimental effect appear to diminish as the duration of the behavior increases. For example, if the data from the first two experimental conditions of

---

[1]In Figure 6 a range of possible outcomes is shown for interval data of each condition. Actually, the probability that the 30-second response will be recorded in four intervals approaches 1.00, and the expected value is virtually the only possible outcome. However, for Figure 6 it has been assumed that there is some finite probability that a 30-second response could occur at the precise moment a 10-second interval begins and be recorded in only three intervals.

Figure 4

Hypothetical results from an ABAB experimental design yielding four different frequencies of a behavior with a duration of 5 seconds. The top section shows the actual frequency of the behavior. The middle section shows the percent of session in which the behavior occurs. The bottom section shows the percent of intervals (or number of intervals out of a 30-minute session) in which the behavior is expected to be recorded. The shaded area is the range of possible values.

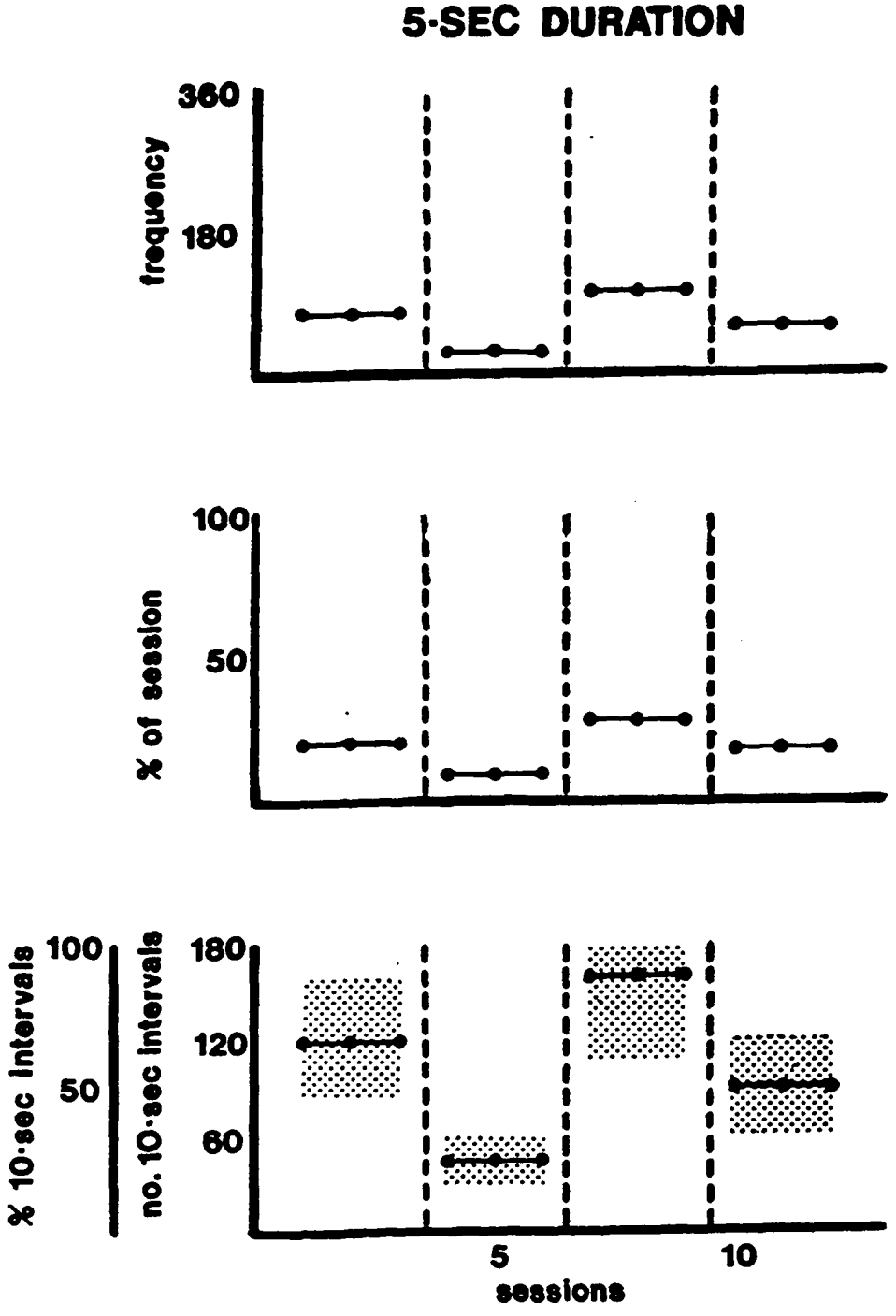


Figure 5

Hypothetical results from an ABAB experimental design yielding four different frequencies of a behavior with a duration of 15 seconds. The top section shows the actual frequency of the behavior. The middle section shows the percent of session in which the behavior occurs. The bottom section shows the percent of intervals (or number of intervals out of a 30-minute session) in which the behavior is expected to be recorded. The shaded area is the range of possible values.

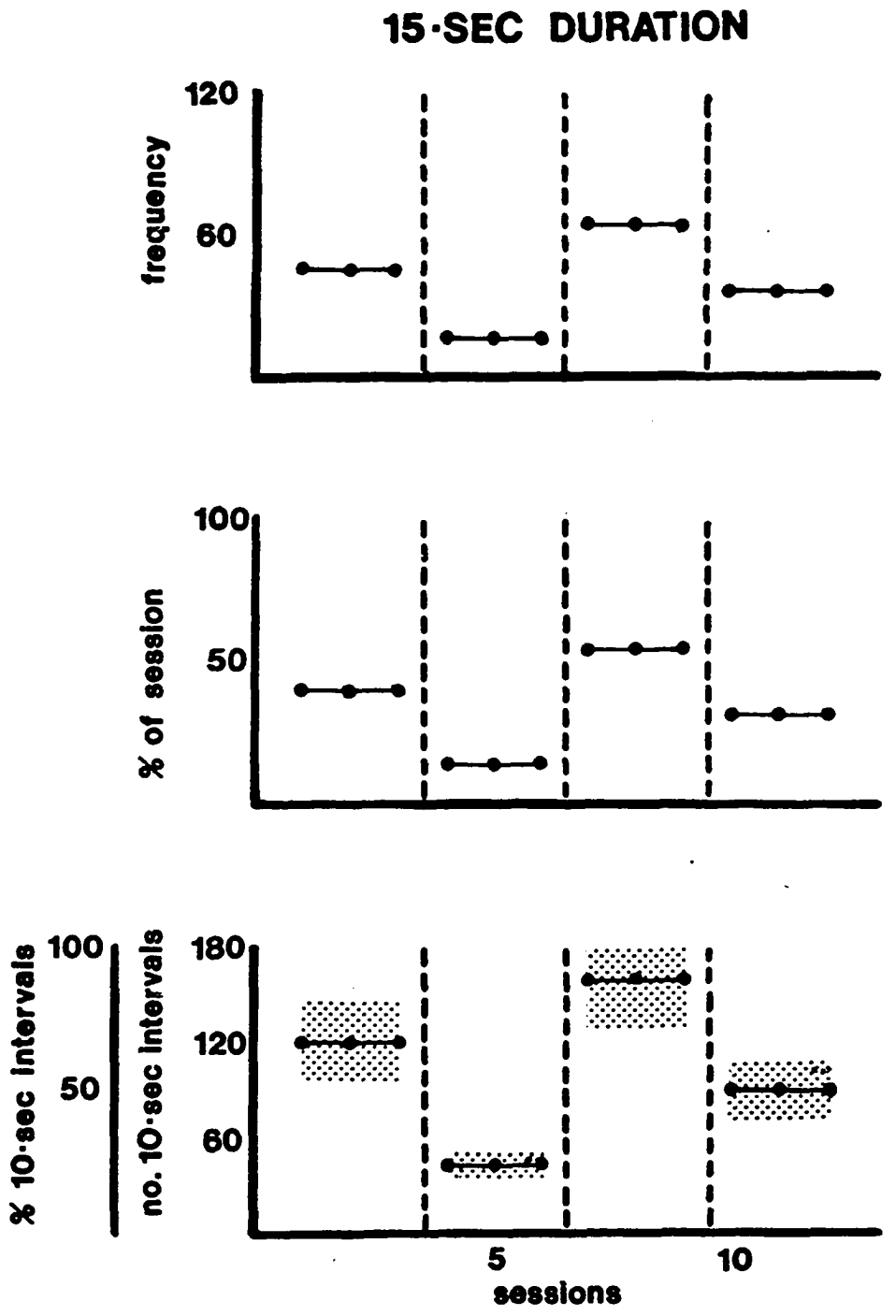
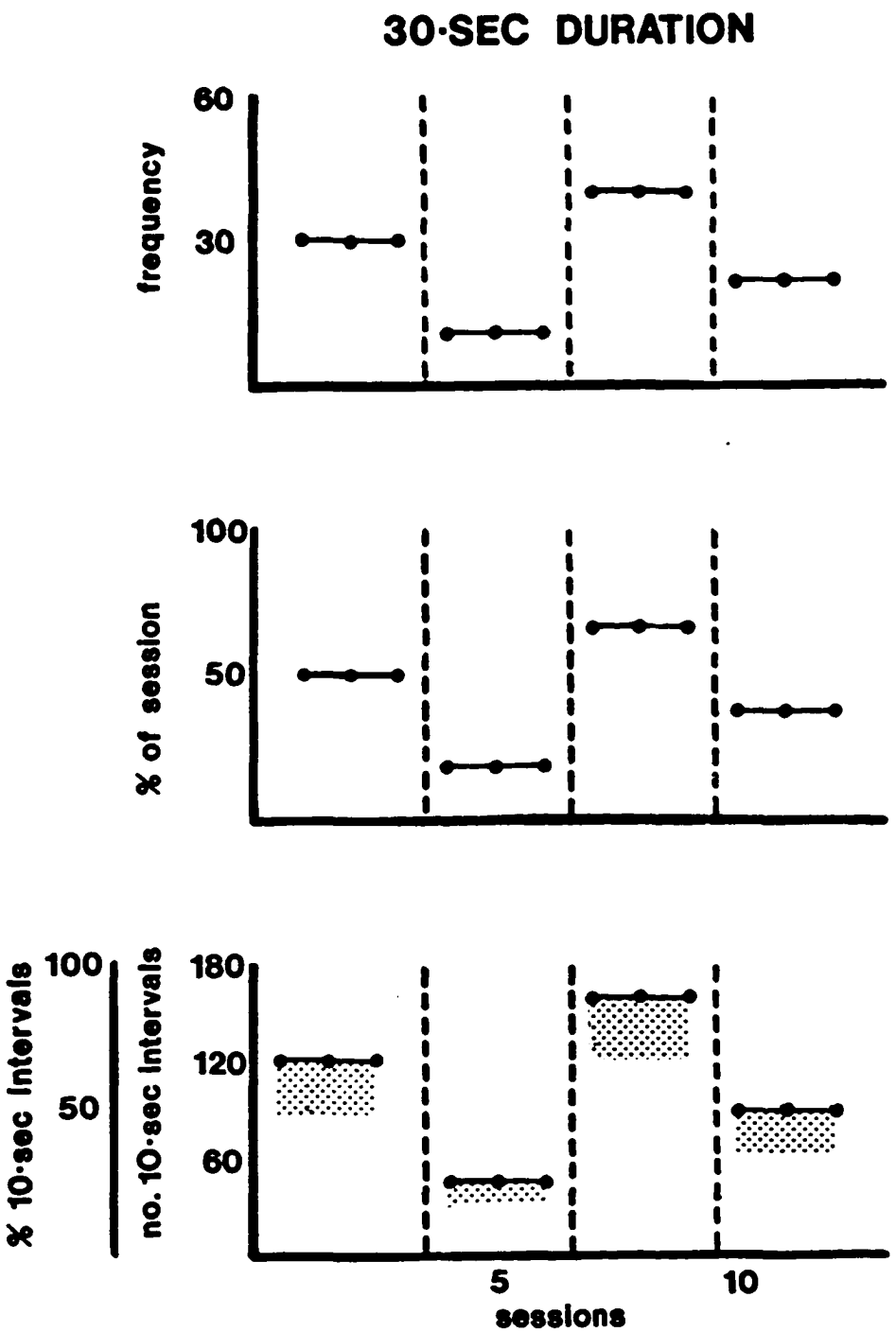


Figure 6

Hypothetical results from an ABAB experimental design yielding four different frequencies of a behavior with a duration of 30 seconds. The top section shows the actual frequency of the behavior. The middle section shows the percent of session in which the behavior occurs. The bottom section shows the percent of intervals (or number of intervals out of a 30-minute session) in which the behavior is expected to be recorded. The shaded area is the range of possible values.



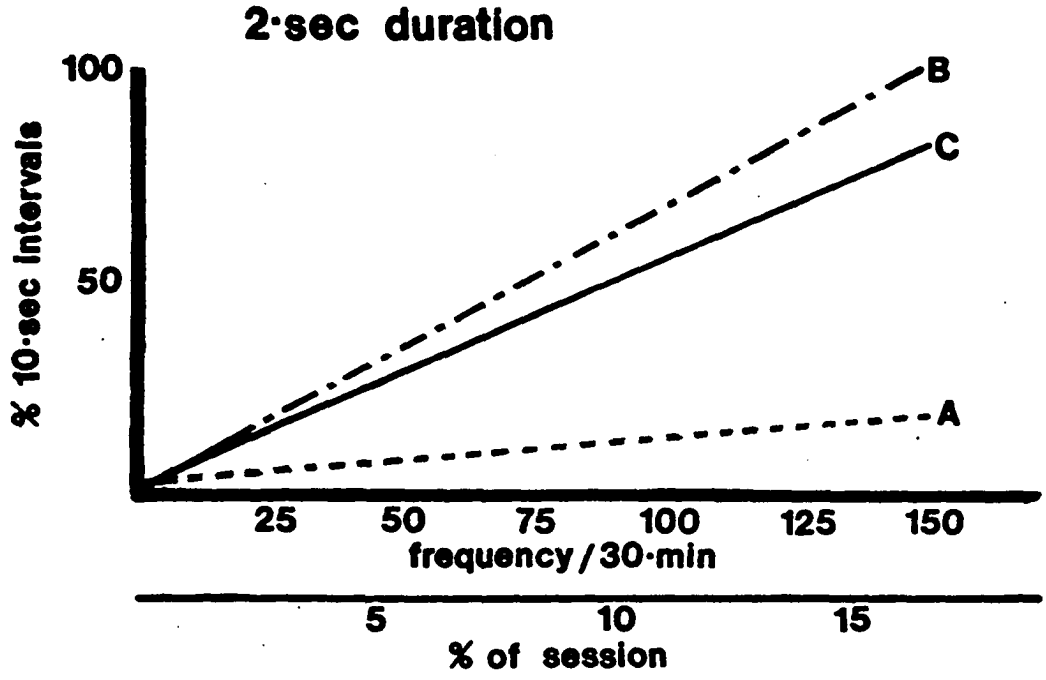
Figures 3 and 5 are compared, it can be seen that the interval data in Figure 5 more closely approximate the true frequency and duration data both in terms of absolute level and in terms of the magnitude of behavior change.

It is not the duration of the behavior, however, which is the determinant of the distortion, but rather the frequency. If a behavior occurs 10 times during a 30-minute session, the distortion in the interval data is the same whether the behavior has a duration of 2 seconds or 30 seconds. This will now be demonstrated. Figures 7 and 8 will illustrate the frequency dependent nature of the distortion resulting from interval recording. The upper and lower sections of each figure represent behaviors with differing durations. Function A in each of these figures represents what would be an ideal relationship between the recorded data and the actual total duration of the behavior being recorded. When the behavior is occurring 10% of the time, it would be recorded in 10% of the intervals; and when the behavior is occurring 33% of the time, it would be recorded in 33% of the intervals. Function B shows the percent of 10-second intervals in which the behavior is expected to be recorded as a function of the frequency or percentage of time the behavior actually occurs. This function is merely a graphic representation of the expected value and is calculated by means of the formula presented earlier. Function B is based on IRTs greater than the length of the interval and represents the mode of a range of possible values. Function C in these figures illus-

Figure 7

Demonstration of frequency-dependent nature of the distortion produced by interval recording

Function A shows the ideal relationship between the actual behavior and the recorded data. Function B is the expected number or percent of intervals in which the behavior is expected to be recorded. Function C is the percent inflation. The upper section shows a behavior with a duration of 2 seconds and the lower section a behavior with a duration of 5 seconds.



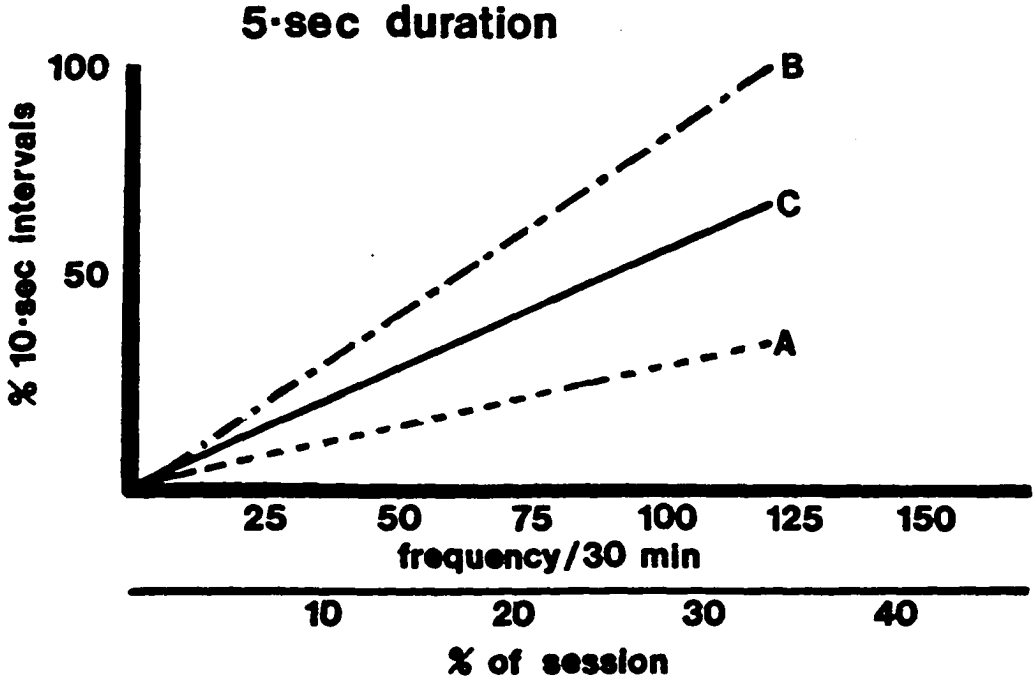
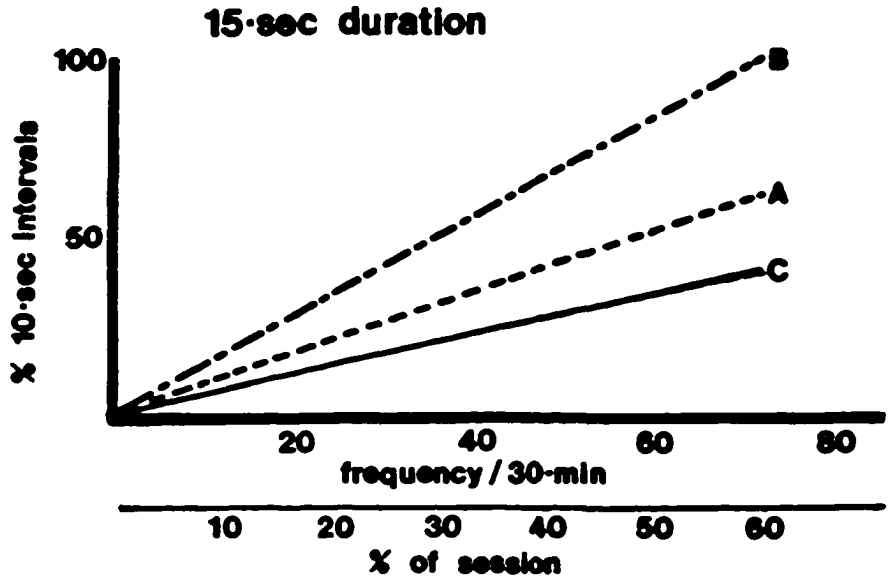
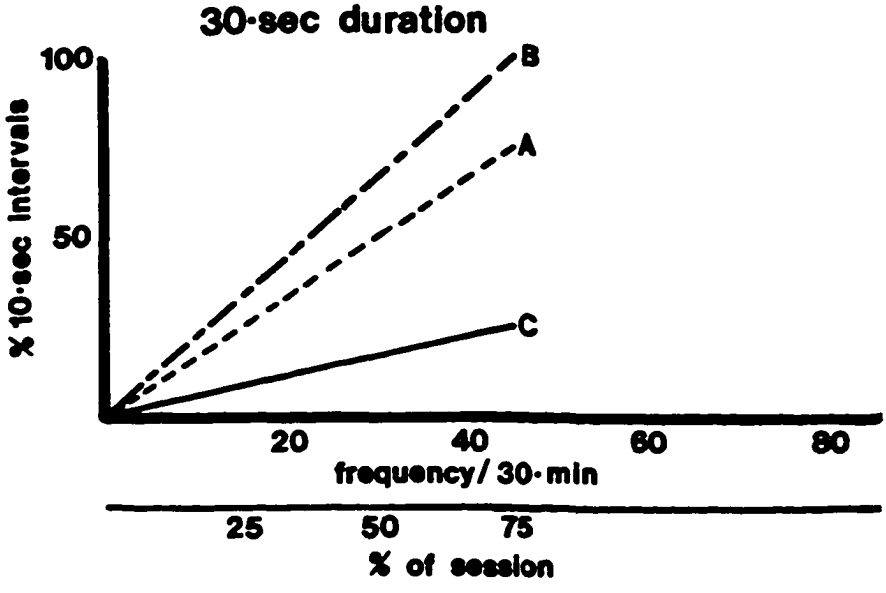


Figure 8

Demonstration of frequency-dependent nature of the distortion produced by interval recording

Function A shows the ideal relationship between the actual behavior and the recorded data. Function B is the expected number or percent of intervals in which the behavior is expected to be recorded. Function C is the percent inflation. The upper section shows a behavior with a duration of 15 seconds and the lower section a behavior with a duration of 30 seconds.





trates the expected percent inflation which results from recording a behavior with the 10-second interval method. This inflation is the difference between the percent of 10-second intervals in which the behavior is expected to be recorded and the actual percent of time the behavior is occurring (function C = function B - function A).[2]

Figures 7 and 8 show that for any particular frequency of the behavior the percent of inflation is the same regardless of the response duration. For example, whenever the frequency of the behavior is 90 per session, 11% inflation is expected, whether the response duration is 2 seconds, 5 seconds, or 30 seconds. Whenever the frequency of the behavior is 40 per session, there is 22% inflation. Thus, if an experimenter were using a procedure which resulted in a change in only the duration of the behavior, and if the IRTs of the responses remain greater than the length of the interval (10 seconds), the use of interval recording would result in the same absolute distortion in all experimental conditions. In such a case, the absolute change in behavior would be accurately represented by the data collected by the interval method. If,

---

[2]It will be noticed that the functions in Figures 7 and 8 are only carried out to certain frequencies. Beyond the frequencies shown on the abscissa the occurrences of the behavior are so numerous that it is no longer possible to have IRTs longer than the interval. In other words, there is a ceiling effect, and any further occurrences of the behavior will result in multiple, unrecorded occurrences of the behavior within an interval. The problem of such short IRTs will be dealt with in more detail later in this paper.

however, the experimental procedure resulted in any change in the frequency of the behavior, a portion of what is reported as behavior change is nothing more than a change in the percent of inflation for differing frequencies of the behavior. The net result would be that there is an exaggeration of the magnitude of the behavior change because the inflation is not equated across the various experimental conditions.

All of the preceding analysis has been based on a comparison between the actual frequency and/or duration of the behavior and the data obtained by the interval method. This comparison has shown that there are several problems with the interval recording method. First, for any given duration of a behavior, interval recording results in differing degrees of distortion depending on the frequency of the behavior. Second, if an experimental procedure brought about a change in the frequency of the behavior, interval recording magnifies that change.

It is important to note that when a behavior is recorded in 100% of the 10-second intervals, it is usually the case that the behavior is not occurring 100% of the time. What is happening is that due to the nature of the recording method, a behavior that takes up even a small portion of the interval is recorded as occurring in the entire interval. In addition, as the behavior increases in frequency, this distortion increases. When IRTs are greater than the interval size it is possible to record the behavior as occurring in 100% of the intervals but the behavior certainly is

not occurring 100% of the time. This is similar to cutting off the top part of the ordinate of a graph and then stretching the remainder of the ordinate so as to maximize the apparent behavior change. This is, in fact, what happens. If one were to compare pre- and post-treatment interval data with the actual pre- and post-treatment levels of the behavior (either frequency and/or duration), the resulting proportions would accurately reflect the actual change in the behavior, provided the data happened to average around the expected value. For example, the behavior change between the first and second phases in Figure 3 shows a change from 100 to 38 in frequency (top section of the figure), from 11.1% of the total session to 4.2% of the total session (middle section of the figure), and from 120 intervals to 45 intervals (bottom section of the figure). Each of these measures is approximately a 38% reduction in the behavior. The interval data, however, look much more impressive than the corresponding frequency or duration data. We would question the honesty of a researcher who took the frequency or duration data from Figure 3, eliminated the higher frequencies or percent of session, stretched out the ordinate and then relabeled the scale from 0% to 100%. When IRTs are greater than the length of the interval used in recording, this is exactly what is happening with interval recording.

A further limitation inherent in interval recording is that it is only within the limited confines of this analysis where variables are held constant or manipulated at will that we can know what the

38% reduction in the recorded data means in comparison to the actual behavior. In the natural environment this reduction could reflect a change in frequency, duration, or IRT of a behavior, or some combination of these. There is no way of knowing which of these factors or combination of factors resulted in the behavior change.

The effect of interval size

As mentioned earlier, Arrington (1939), Bijou et al. (1969), and Mattos (1971) have all recommended the use of smaller intervals when the behaviors being recorded are either of short duration or high frequency. The argument for adjusting the interval size is that smaller intervals increase the ability of the researcher to discriminate between such behaviors. If the smaller intervals allow the researcher to make such discriminations, it should reduce the distortion that was made evident in the preceding section. The accuracy of this argument can be assessed by comparing the validity of interval data obtained by recording with different sized intervals.

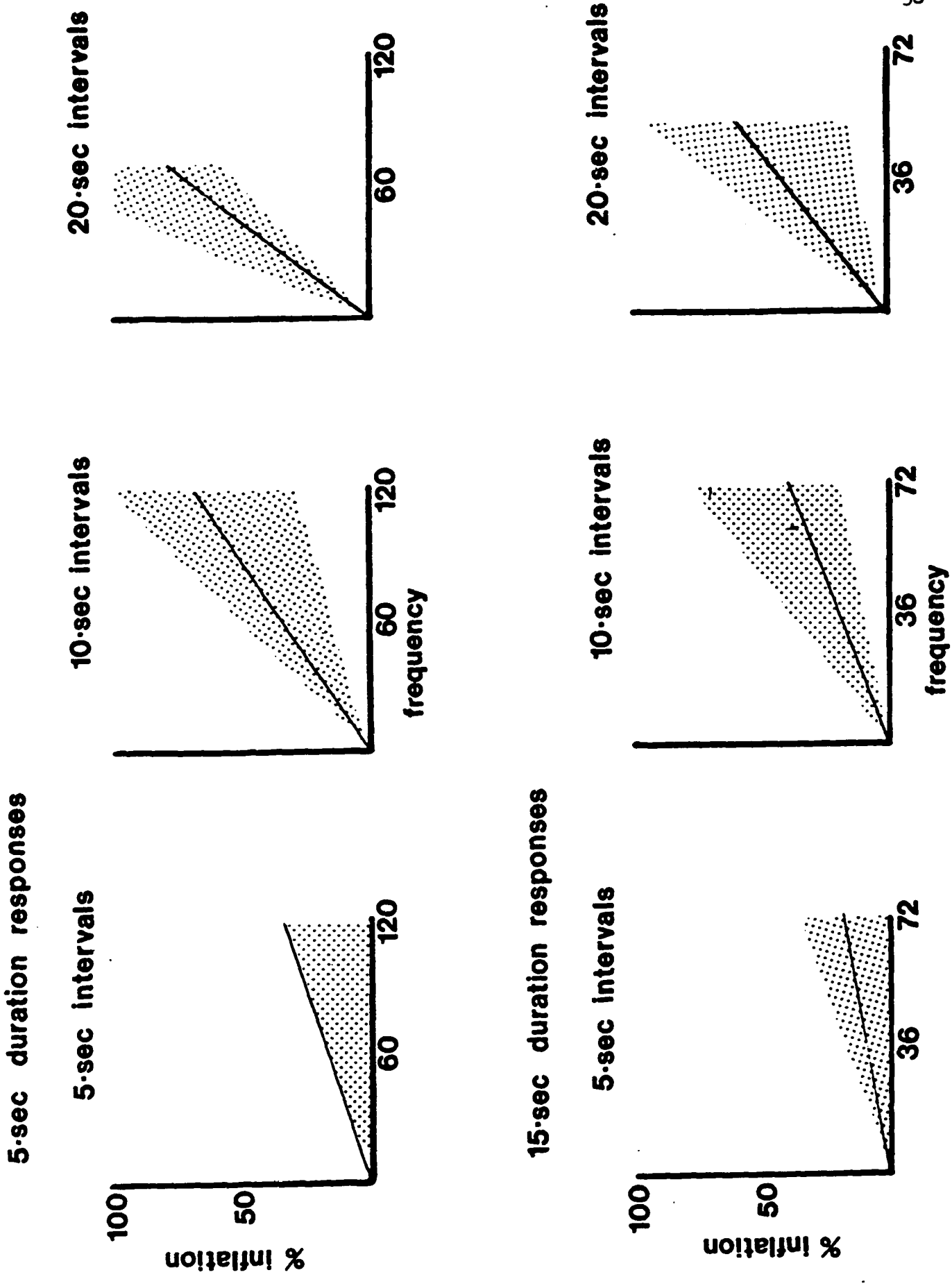
As was shown earlier, the most probable data, or expected value, can be calculated for any particular combination of response duration and frequency, given a particular size of recording interval. Using this formula the expected value was calculated for two hypothetical responses--a 5-second response and a 15-second response--under three recording methods: 5-second intervals, 10-sec-

ond intervals, and 20-second intervals. The difference between the expected value, expressed as percent of intervals, and the actual duration of responding, expressed as percent of session, was calculated. This difference always constituted an inflation of the behavior (because IRTs were held larger than interval size, thus preventing multiple responses within an interval), and these expected inflations or distortions due to interval recording are presented in Figure 9 for each frequency of both the 5-second and the 15-second duration response (up to the maximum, given the duration and IRT). Figure 9 should not be interpreted to mean that proportional changes in the frequency of a behavior are inaccurately reflected in interval recorded data. As discussed earlier, interval data accurately reflect the proportional changes (e.g., 150% increase, or 70% decrease) in the frequency and/or duration of a response.

The inflation is the difference between the recorded data and the actual frequency and/or duration of the behavior. Also shown in Figure 9 is the range of possible inflation, represented by the shaded area of each graph. As can be seen, for the frequencies and durations illustrated, the use of 10-second intervals results in twice the percent inflation that would result from the use of 5-second intervals. If 20-second intervals were used to record the behavior, the inflation would be twice as great as that which would result from the use of 10-second intervals and 4 times as great as 5-second intervals. The longer the interval used to record the

Figure 9

The percent absolute inflation as a function of the duration of a behavior (5 seconds and 15 seconds) and the size of the interval used to record the behavior (5-, 10-, and 20-second intervals). The shaded area is the range of possible inflation.



behavior in question, the greater will be the resulting inflation in the recorded data. Even with intervals as short as 5-seconds, the inflation inherent in the method of data collection will distort the absolute results obtained under any particular condition. However, the distortion is reduced by the same factor that the experimenter reduces the interval size.

The effect of interresponse time

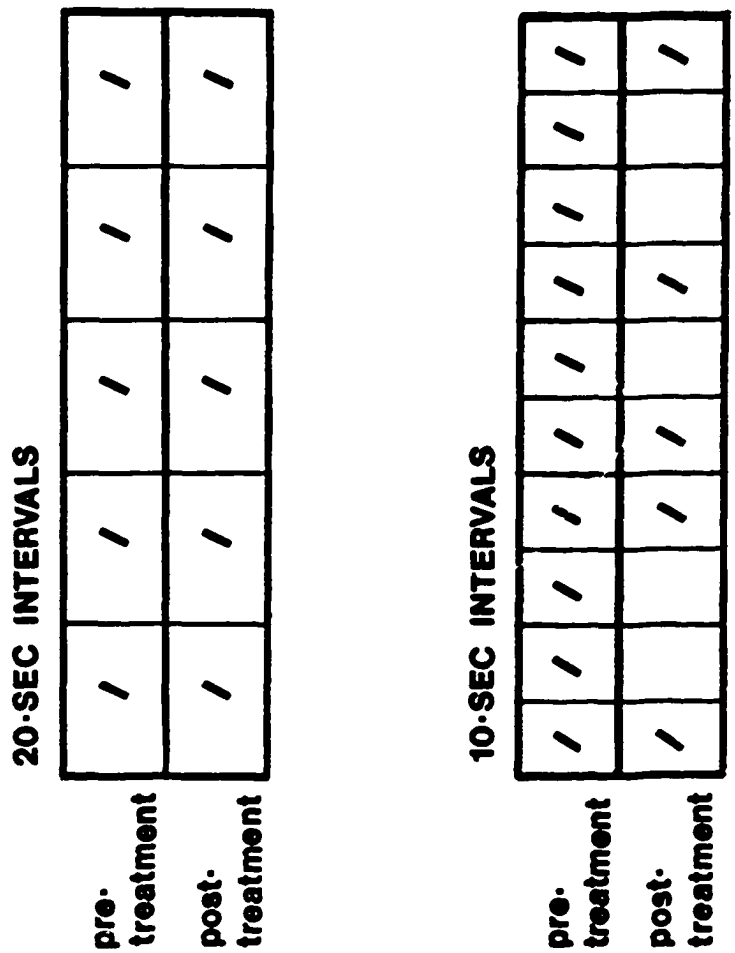
Up to this point in the analysis the IRT has been held equal to or greater than the length of the interval. Some of the problems of IRTs shorter than the length of the interval will now be dealt with. When the IRTs are shorter than the length of the interval and the duration of the behavior is short it is possible for multiple occurrences of the behavior to take place in any one interval. It might be argued that such a pattern of behavior would reduce the inflation previously shown to result. It is true that the amount of inflation is reduced, but the researcher is merely substituting one source of error--unrecorded occurrences of the behavior--for another source of error--the inherent inflation of the recording method. Because the behavior analyst is not able to discriminate the multiple occurrences of the behavior at this high rate (short IRT), it is possible that what might be considered an effective treatment program would go unnoticed. For example, a behavior which occurred two or three times during every interval would only be recorded once. If after a treatment was introduced

it occurred only once during every interval, it would still be recorded as occurring only once. In spite of a sizeable reduction in the rate of the behavior, interval recording has masked the behavior change.

Short IRTs also interact with interval size and can mask behavior change. It was shown earlier that the shorter the interval used to record the behavior, the less the inflation of interval recording. It is also the case that the shorter the interval, the less likely there are to be multiple occurrences of the behavior within an interval. For example, Figure 10 is a portion of a data sheet with the slash marks in each interval representing the occurrence of a behavior. The upper section of Figure 10 is the pre-treatment and the post-treatment data recorded by using 20-second intervals. Both pre- and post-treatment data show the behavior occurring in 100% of the intervals. If 10-second intervals had been used to record the same behavior, the data sheet might have looked something like the lower section of Figure 10. Here the pre-treatment data show the behavior occurring in 100% of the intervals, but the post-treatment data show the behavior occurring in 50% of the intervals. In this case, the use of 20-second intervals has interacted with short IRTs to mask what could be considered an effective treatment procedure. Thus, problems with interval recording are the greatest with longer intervals where multiple occurrences of a behavior with short IRTs causes a deflation or masking of the actual strength of a behavior.

Figure 10

Sample data sheets showing pre- and post-treatment records when 20-second or 10-second intervals are used to record the behavior.



Empirical Analysis of Interval Recording

Perhaps an example or two of actual instances of distortion will demonstrate some of the practical problems in the use of interval recording. Classroom behavior was collected using the Datamyte system.[3] With this instrument the initiation and cessation of behaviors are entered on a keyboard similar to that of an adding machine. As the behavior codes are entered on the keyboard, the codes and the time (to the nearest second) are recorded on a cassette tape. Subsequent computer analysis of the cassette tape furnishes the researcher with much the same information as an event record; namely frequency, duration, IRT, and temporal distribution of the responses.

The behavior recorded in the following examples was out-of-seat and the recording sessions were 30 minutes long. Out-of-seat behavior was defined as "any time the seat of the pants of the subject was not in contact with the seat of the chair." In the first example, the behavior occurred 7 times with a total duration of 405 seconds. Thus, the behavior was actually occurring 22.5% of the session. If the behavior is recoded into 10-second intervals, the interval method would show that the behavior occurred in 51 or 28.3% of the intervals. In this particular instance, there is an

---

[3]Manufactured by Electro/General Corporation, 128 Jackson Avenue North, Hopkins, Minnesota 55343.

overestimation of 5.8% in the total duration of the behavior. And, if one were looking at the interval data as a reflection of the frequency of the behavior, the distortion would be even greater (an actual frequency of 7 compared to an interval frequency of 51).

In another example, out-of-seat behavior occurred 50 times during the 30-minute session for a total duration of 976 seconds which represents 48.7% of the total session. However, when the behavior was recoded into 10-second intervals, it was found to occur in 111 intervals which represents 61.7% of the session. Here, interval recording resulted in a 13% inflation in the total duration of the behavior.

A final example also points out the distortion resulting from the use of interval recording. Out-of-seat behavior was recorded 35 times during a 30-minute session with a total duration of 331 seconds or 18.3% of the session. After recoding into ten second intervals, the behavior occurs in 71 or 39.4% of the intervals. A 21.1% inflation of the total duration of the behavior has resulted.

Recording actual behavior in a classroom setting and then recoding the data into 10-second intervals demonstrates that interval recording does result in distortion and this distortion is usually inflation. Even though there were occasions of multiple occurrences of the behavior within an interval in the above examples (especially the second one), the inherent inflation of interval recording predominates in every instance. The examples were not carefully chosen to demonstrate distortion; they happen to represent data from the

first three days of an experiment to modify out-of-seat behavior. In all three instances the distortion is substantial and could lead one to question the internal validity of any data collected by the interval recording method.

DISCUSSION

This analysis has touched on some of the variables affecting the accuracy of the interval recording method. All behaviors, whether in the world of the behavior analyst or the theoretical world of this paper, exhibit the very same properties upon which this analysis is based, i.e., frequency, duration, and IRT. To the extent that real behaviors exhibit the same properties, the analysis applies to those behaviors. But it is only within the confines of this paper that the interaction of rate, duration, and IRT can be controlled and their effect on the recorded data observed. In the natural environment these variables change in unmeasured ways. Thus, the researcher has no knowledge as to which of these three variables was responsible for the reported behavior change.

The problems with interval recording are multiple. First, there is the distortion in the absolute level of the behavior due to either inflation, as was shown by use of the expected value formula, or deflation, due to unrecorded multiple occurrences of the behavior. Second, the distortion is a function of frequency, duration, IRT, and interval size. When recording behaviors in the natural environment these variables interact and the amount of distortion or the nature of the distortion (inflation or deflation) is not known. Third, interval recording is not an accurate method of estimating either the frequency or the duration of a behavior. Multiple occurrences of a behavior within an interval make interval

recording an unacceptable estimator of frequency and crediting an entire interval when the behavior takes up only a small portion of the interval makes interval recording unacceptable as an estimate of duration.

Baker and Whitehead (1972) make note of the inflation resulting from interval recording and recommend that an event recorder be used. Such an instrument not only measures the occurrence of an event, but its duration as well. The event recorder can also be used to record several behaviors simultaneously and give an accurate representation of these behaviors. The Datamyte is an example of one type of event recorder which may be used.

However, due to the cost of event recorders and their foreign nature in classrooms, homes, and the like, this is often not a practical alternative. In such a case, a form of interval recording may be useful despite its limitations. Increased accuracy in recording would result if only initiations of the behavior were recorded and multiple initiations within an interval were also recorded. Thus, the interval method would really be reduced to frequency recording, but a record of the temporal pattern of the behavior would be made at the same time. The exaggerated magnitude of behavior change resulting from interval recording would be eliminated and it could be said with increased confidence that "Yes, the experimental treatments did make a difference in this specific experimental instance."

REFERENCES

Arrington, R. C. Time sampling studies of child behavior. Psychological Monographs, 1939, 51, (2).

Baer, D. M., Wolf, M. M., and Risley, T. R. Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1968, 1, 91-97.

Baker, J. G., and Whitehead, C. A portable recording apparatus for rating behavior in free-operant situations. Journal of Applied Behavior Analysis, 1972, 5, 191-192.

Bijou, S. W., Peterson, R. F., Harris, F. R., Allen, K. E., and Johnston, R. S. Methodology for experimental studies of young children in natural settings. The Psychological Record, 1969, 12, 177-210.

Campbell, D. T., and Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Hawkins, R. P., and Dotson, V. Reliability scores that delude: an Alice in Wonderland trip through the midleading characteristics of inter-observer agreement scores in interval recording. Paper read at Third Annual Symposium on Behavior Analysis in Education, Lawrence, Kansas, 1972.

Hays, W. L., and Winkler, R. L. Statistics: probability inference and decision Volume 1. New York: Holt, Reinhart, & Winston, 1970.

Mattos, R. L. Some relevant dimensions of interval recording. Academic Therapy, 1971, 6, 235-244.

O'Leary, K. D., Becher, W. C., Evans, M. B., and Saudargas, R. A. A token reinforcement program in a public school: a replication and systematic analysis. Journal of Applied Behavior Analysis, 1969, 2, 3-13.

Reid, J. B. Reliability assessment of observation data: a possible methodological problem. Child Development, 1970, 41, 1143-1150.

Romanczyk, R. G., Kent, R. N., Diament, C., and O'Leary, K. D. Measuring the reliability of observational data: a reactive process. Journal of Applied Behavior Analysis, 1973, 6, 175-184.

Worthy, R. C. A miniature, portable timer and audible signal-generating device. Journal of Applied Behavior Analysis, 1968, 1, 159-160.