SAMPLING OF SPEECH PATHOLOGY TREATMENT ACTIVITIES: AN EVALUATION OF MOMENTARY AND INTERVAL SAMPLING PROCEDURES

R. H. BROOKSHIRE, L. S. NICHOLAS, *and* K. KRUEGER

Veterans Administration Hospital, Minneapolis, Minnesota

Videotaped samples of aphasia treatment sessions were coded, using the Clinical Interaction Analysis System (CIAS), a 39-category system for recording the events that occur in clinician-patient interactions during aphasia treatment sessions. These coded records were then sampled according to various schedules and procedures and the fidelity with which each sampling schedule and procedure represented the content of the entire treatment record was evaluated. In addition, trained observers coded videotaped samples of treatment, using the CIAS with a number of sampling schedules and procedures. The fidelity with which these observers' records represented the content of the treatment sessions sampled was then evaluated. The results of the analysis indicated that momentary sampling at intervals distributed throughout the session generates more accurate records of session content than single longer samples taken from the session, unless those single samples comprise a major part of the session, and that sampling representativeness remains high even when only one event in ten is sampled, if sampled events are uniformly distributed throughout the session.

The need for systems for quantifying, describing, and recording events which occur within treatment activities conducted by speech pathologists has been recognized for a number of years (Johnson, 1969; Boone and Prescott, 1972; Schubert, Miner and Till, 1973). Such quantification and recording systems are useful (and perhaps crucial) in clinical supervision, in construction and evaluation of clinical treatment programs, and in research designed to evaluate the effectiveness and efficiency of various treatment techniques.

During the 1950's and 1960's a number of systems for coding and recording behavioral events which occur within interpersonal interactions were developed. Bales (1950, 1970) provided one of the first formalized systems for quantification of interpersonal interaction in communication situations, and Flanders (1960, 1963, 1970) has extended Bales' system to the analysis of teacher-student interactions. A number of systems for analysis of clinician-patient interactions in speech pathology treatment contexts have also been devised (Johnson, 1969; Boone and Prescott, 1974; Schubert, et al, 1973; Mowrer, Baker and Owen, 1968; and Diedrich, 1973). In general, these are multiple-category systems by which events which occur within clinician-patient interactions may be identified, labeled, and recorded. The frequencies

of various events occuring within treatment sessions can then be tabulated, and descriptions of the content of the sessions can be formulated from these frequency data.

Most of the interaction analysis systems published during the last twenty years use a substantial number of categories to which events are assigned. Bales' (1970) system used 12 categories, Boone and Prescott's (1972) system utilizes 10, and their (1974) extension of that system expanded it to 19. Johnson's system (1969) contains 40 categories. It is generally the case that systems with larger numbers of categories are more sensitive than those with fewer if the categories can be reliably coded, explicitly defined, and are minimally redundant. In order to obtain an accurate and detailed description of an interaction, one should record as much of the interaction as possible. However, as the complexity of the recording system increases, its suitability for live, on-line recording of the interaction diminishes because observers cannot keep up with the flow of events when they must make decisions about partitioning those events among a large number of categories. Therefore, almost all interaction analysis systems make compromises between the limited capacity of observers, the multidimensionality of the interaction recording system, and the rapid rate at which interaction events to be recorded occur.

Because of these considerations, some systems use audio or videotape recordings, or typed transcriptions of interactions (usually made from audio or videotape recordings), and allow the observer to replay or re-examine any part of the recorded or transcribed interaction as often as necessary to categorize the events. These systems are sometimes called off-line systems, because the observer is not required to make decisions in real time, but has a permanent record of the interaction available, and can take as much time as needed, looking at, or listening to, the events as many times as necessary to make the appropriate coding decisions.

Other systems require the observer to record events as they occur, "on-line." Most systems for on-line recording permit the observer to code only a portion of all the events that occur, by allowing him to sample events at given time intervals, or by allowing him to code only those events that he can under the constraints of real time.

Some systems instruct the observer to watch the interactions until an "event" occurs and to record and make decisions about that event and ignore others. When the event has been entered, the observer looks for the next one, and so forth (Kunze, 1967). Most systems use a technique called "time-sampling," in which the presence or absence of events at particular points in time is evaluated and tabulated. There are four basic time-sampling techniques (Powell, Martindale and Kulp, 1975; Repp, et al, 1976). In *whole-interval time-sampling*, the observer decides, at the end of each of a series of predetermined intervals, whether a given event or behavior has been present throughout the interval. If it has, the observer scores that interval as one in which the event or behavior occurred. In *partial-interval time-sampling*, if the event or behavior is present during any part of the interval, that interval is scored as one

in which the behavior or event occurred. In *momentary time-sampling*, the observer makes assessments at specified moments (regular or irregular). If the event or behavior is present at the moment of observation, its presence is recorded. In *frequency recording*, events or behaviors are counted during relatively large segments of the session (or over the entire session) and the data are summarized as events per unit of time, events per sampling interval, or events per session. In all time-sampling procedures except frequency recording, multiple occurrences of an event or events within an interval (or at the moment of observation) are not coded; the presence of the event within the interval is recorded, not its frequency within the interval. By increasing the length of the interval in time-sampling procedures, one can make it possible for coders to use recording systems of greater length and complexity, because once the occurrence of an event category is recorded, the observer need not be concerned with that event category until the next sampling interval begins. The most popular sampling interval is the 3 sec interval proposed by Bales (1950), used also by Flanders (1970), and by those who have developed systems based on their models (Schubert, et al, 1973). Ten-second interval sampling is also frequently used, especially by operant investigators (Thomson, Holmberg and Baer, 1974).

A third technique for on-line coding is to have observers record every codable event that occurs in the interaction. On-line recording of all interaction events usually requires that the coding and recording system be extremely simple, or that the observer be given some means for rapidly recording the events which occur. A few systems require coders to record all codable events on-line, using paper and pencil data recording, although it is questionable whether users of these systems actually record all codable events, or do so reliably (Fine and Zimet, 1956). Some systems provide the observers with electronic or electromechanical keyboard data-accumulation devices which allow the observation and recording of relatively large amounts of information on-line, once observers are trained in the use of the devices (Diedrich, 1976).

It is not unusual for coders to spend five to ten times more time coding an interaction from a recording than the interaction took in real time. If observers work from transcripts of sessions (Cromwell and Scheidel, 1961), then the time needed to prepare the transcript must also be considered when one evaluates the time investment required by the system.

Finally, in the case of both on and off-line coding, the recorded data must be tabulated, summarized, and interpreted. The amount of time required for these activities is influenced by the extent and efficiency of the data tabulating techniques, and the amount of data collected. It is necessary, therefore, to examine ways to minimize time investment in observational recording. This can be accomplished by sampling, rather than exhaustively recording events, without compromising the validity or the reliability of the measures obtained.

Literature on the validity and accuracy of various methods of sampling in observational recording is surprisingly limited, given the wide use of various sampling procedures. Powell et al (1975) and Powell et al (1977) have

demonstrated that when continuous events, such as "time spent working" are recorded using time-sampling procedures, partial-interval time-sampling tends to overestimate the proportion of time that behaviors are actually present, while whole-interval time-sampling tends to underestimate the proportion of time the behaviors are present. They conclude that momentary time-sampling is more accurate and more easily accomplished, and should be used in investigations where duration is the response dimension of interest. They also found that, as the sampling interval increased, the amount of error attributable to time-sampling also increased.

Thomson, et al (1974) evaluated three patterns of 10-sec (momentary) time–sampling distributions: *contiguous*–16 min of continuous time sampling from a 64-min session; *alternating*–alternate 4–min intervals of time–sampling during the first or last half of a 64–min session; and *sequential*–4-min *sampled* every 16 min (12 min between sample intervals) throughout a 64–min session. The records generated by each sampling distribution were compared with records generated by exhaustive recording of the entire 64 minutes. Contiguous sampling (16 min of continuous sampling) generated the greatest (approximately 39%) error. Alternating sampling (alternate 4 min intervals in one–half of session) also generated substantial (33%) error. Sequential sampling (4 of every 16 min) generated substantially lower error (approximately 10%) than either of the other sampling techniques. The authors concluded that it is better to sample briefly but repetitively throughout a session (the sequential method) than to sample over longer, unbroken intervals which represent only part of the session.

Repp et al (1976) compared the effectiveness of momentary time-sampling partial-interval time-sampling, and frequency recording in representing the actual frequencies of events generated by electromechanical equipment according to predetermined rates and patterns throughout 180-min sessions. They found that time-sampling ". . . does not provide data that properly represents events in the environment, and some conclusions from studies using this data collection method are clearly in question. . . ." They found that frequency recording (of the entire session) provided a more accurate representation of what occurred in the session than either time-sampling procedure, with time-sampling generating greater error as event rates increased.

Olsen (1972)[1] compared the effectiveness of various sampling durations in representing the content of speech pathology treatment sessions, and concluded that a 5–min sample taken from the middle 20 min of 30–minute treatment sessions is representative of the session as a whole. Schubert and Laird (1975), using a 12–category system, compared the representativeness of 3–min samples taken from 15 minute articulation therapy sessions. They concluded that ". . . a three–minute period was sufficient to obtain a representative

[1]Comparisons of sequential interaction patterns in therapy of experienced and inexperienced clinicians in the parameters of articulation, delayed language, and voice disorders. Unpublished doctoral dissertation, Univ. of Denver (1972).

sample of clinician-client interaction during a therapy session. . . ." They noted, however, that some behaviors did not stabilize until the second three minutes of the session, suggesting that the first 3-min sample might not be as representative of the session as later samples would be.

The literature suggests that the various methods of interval sampling may not provide accurate representations of the frequency or duration of behaviors or events observed over time, especially if the events occur at high rates. There is evidence that frequency recording, either at intervals within the session or throughout the session, may provide more representative measures than interval sampling. Several investigators have suggested that sampling should be distributed throughout the session (Bijou, Peterson, and Ault, 1968; Thomson, et al, 1974; Powell, et al, 1975), although Schubert and Laird (1975) and Olsen (1972) have concluded that a single continuous three or five-minute sample taken from 15 or 30 minute treatment sessions provides a representative sample of the entire session.

As part of our development of an interaction analysis system (Brookshire, 1976), an extensive investigation of the representativeness of various sampling procedures was conducted. The results suggest that when sampling treatment sessions as opposed to observing and recording them in total, certain general principles must be observed to insure accurate representation of the interaction.

METHOD

Production of Treatment Session Master Records

The Clinical Interaction Analysis System (CIAS) (Brookshire, 1976) was used to code and record events which occurred in clinician-patient interactions within a number of 10-minute videotaped samples from aphasia treatment sessions, which had been used in establishing the validity and reliability of the CIAS. The CIAS employs 39 event categories to record interaction events (Appendix).

As part of the development of the CIAS, 40 videotaped samples of aphasia treatment were gathered from aphasia treatment facilities throughout the United States. From each of these tapes, a 10-min segment of treatment was selected (its location was randomly determined) and coded, using the CIAS, to develop a master coding record for each sample. Then, the proportional occurrence of each of the 39 event categories was computed for each of the 40 tapes, by dividing the frequency-of-occurrence for each event category within a tape by the total number of events contained within that tape. The master record for each 10-min videotaped sample was prepared by two observers who had demonstrated their reliability (category-by-category percentage agreement > 0.90) in the use of the CIAS. These two observers viewed each 10–min tape, and coded it completely using the 39 event categories shown in the appendix, rewinding and repeatedly viewing given events

as necessary to arrive at an agreed-upon code for each event in the tape. The judgments of a third observer, also with previously established reliability with the coding system, were used to resolve any disagreements between the other two observers. In this manner, a complete and reliable record of the events which occurred in each 10-min videotape was generated. These records were used in the analyses of sampling procedures which follow.

Momentary Sampling: Exact Ratio

Procedures. If an observer is instructed to code as much as possible without recourse to video or audio recordings or transcriptions of treatment sessions, and if the complexity of the coding system and the rate of events in the session preclude the coding of every event which occurs, then we might expect that the observer would record some proportion of total events, with recorded events distributed relatively evenly throughout the session. To determine the effects of this method of sampling, we first evaluated an idealized version of such sampling, and subsequently checked the results of this version against the performance of observers in real life. In order to provide a basis for judgments regarding the acceptability of various sampling procedures, we established the criterion that 90% of event categories sampled by a sampling schedule must be within 10% (absolute) of the actual frequency of occurrence for each event category sampled.

To construct an idealized prototype of the data that observers might generate if they were instructed to code as much of sessions as they could without recourse to recordings, we sampled from the master records of 5 ten-minute aphasia treatment sessions according to seven sampling schedules; every other event (50% sampling), every third event (33.3% sampling), every fourth event (25%), every fifth (20%), sixth (16.7%), eighth (12.5%) and every tenth event (10% sampling). The five ten-minute tapes contained 548 total events. The number of events contained within individual tapes ranged from 81 to 167 (mean = 110 events). The proportions of occurrence (event category: total events) of each of the 39 categories on each tape were calculated for the records generated by sampling at each of the seven sampling schedules. Then, for each of the 39 event categories, the proportion occurrence generated by each sampling schedule was subtracted from the proportion occurrence on the master record for the tape. This generated a difference score (the difference between the master record and the sampling schedule) for each of the 39 event categories sampled on each of seven schedules. These difference scores were then summed (ignoring minus signs, to avoid cancelation of positive sampling error in one category by negative sampling error in another) and divided by 39 (categories) to produce a mean absolute error proportion across all 39 categories for each sampling schedule.

Results. The solid black dots in Figure 1 represent the results of the calculations described in the preceding paragraph. The mean differences from the master record and the standard deviations of those differences for each of the

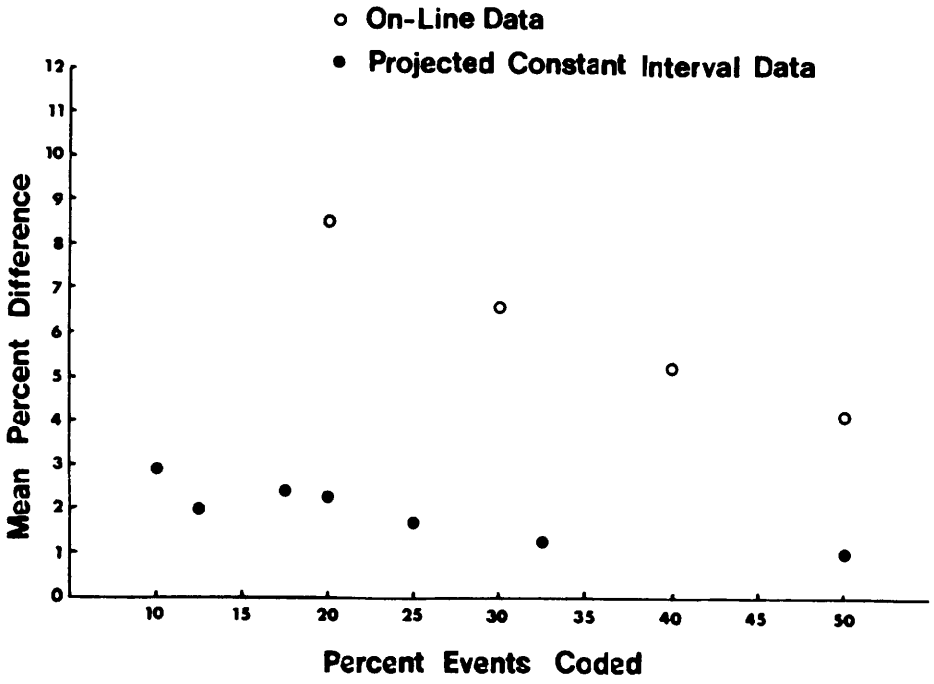

FIGURE 1. Mean percent difference from true proportions of events within treatment sessions for idealized (Projected Constant Interval Data) and real (On-Line Data) observers sampling treatment sessions on various momentary sampling schedules.

five tapes at 50%, 20% and 10% sampling schedules are presented in Table 1. It can be seen in Figure 1 and Table 1 that, even when only every tenth event was sampled, error introduced by these sampling schedules was negligible. Table 1 also shows that the standard deviations of the sampling error gen-

TABLE 1. Means, standard deviations, and 90% confidence intervals of absolute differences between master record and event samples from that record with 50%, 20%, and 10% event sampling schedules.

	Tape #					
Sampling Schedule	*1*	*2*	*3*	*4*	*5*	*90% CI*
50%						
Mean (%)	0.93	0.85	0.59	1.26	1.89	0-5.20
SD	2.10	1.76	1.03	2.72	4.84	–
20%						
Mean (%)	2.13	2.48	1.88	2.51	2.23	0-9.08
SD	3.58	5.71	3.45	3.95	4.01	–
10%						
Mean (%)	2.48	2.68	1.86	3.98	3.48	0-10.50
SD	3.63	4.67	3.97	5.27	5.49	–

CI upper limit = Mean + 1.65 *SD*.
CI lower limit = zero, because absolute differences were used.

erated by the various sampling schedules were small, suggesting that the mean error percentages were representative of the 195 (5 tapes × 39 categories) individual difference scores obtained. As predictions of tape-by-tape sampling representativeness which might be obtained using these sampling schedules, 90% confidence intervals (using areas under the normal curve) were calculated for each schedule. The limits for these 90% confidence intervals are presented in Table 1. The results of these calculations suggest that even when only one in ten events is sampled, the records for all event categories sampled will be within 10.5% (absolute) of the true proportions of those event categories within the entire treatment session (91% of categories on a 1/10 schedule will meet the 10% error criterion described earlier).

Momentary Sampling: Real Observers

Procedures. In order to determine whether real observers who sampled events on equivalent schedules would also generate records with low error, we examined the representativeness of records generated by real observers who adventitiously generated sampling schedules similar to the idealized schedules previously evaluated. Five trained, reliable observers coded the events which occurred in five 10-minute videotaped samples of aphasia treatment, using the CIAS. The observers played the tapes without stopping or rewinding, and were instructed to use the CIAS to record as many events from each tape as they could, ignoring those events which occurred while they were coding or recording a previous event. The percent of total events in each tape coded by each observer was then calculated.

Results. The percentage range of events coded was 23-86%, and the mean percent events coded for all observers across all tapes was 40.2%. The records in which observers coded 20 ± 3% (N = 5), 30 ± 3% (N = 6), 40 ± 5% (N = 8), or 50 ± 5% (N = 5) of any session were then evaluated to determine the fidelity with which these records represented the actual frequency of events in a session. Absolute difference scores were computed in the same manner as for the idealized sampling schedules by subtracting event category proportions coded by each judge from the master record proportions, as described previously. The results of these computations are presented in Figure 1 (open circles). Although the sampling error for real observers was somewhat greater than for comparable idealized sampling schedules, there was still less than 10% absolute error of estimation by real observers, even when, on the average, only one event in five was sampled.

It seemed important to know whether observers who coded 20% of all events tended to code every fifth event, whether those who coded 30% of all events tended to code every third event, and so forth, because it was possible that observers could "clump" coded events, and thus depart from the assumptions of the idealized schedule described earlier. Therefore, we computed the number of events intervening between coded events for observers who coded 20%, 30% or 50% of events (Figure 2). Figure 2 shows that there is a strong

tendency for all distributions to be skewed toward the left, so that even observers who recorded only 20% of all events sometimes coded every third event, in some cases coded every other event, and occasionally coded consecutive events. If our observers had coded clumps of events, separated by clumps of uncoded events, the distributions in Figure 2 would be bimodal, with large percentages of zero values and large percentages of large values. The distributions, though skewed, are unimodal, suggesting that clumping did not occur.

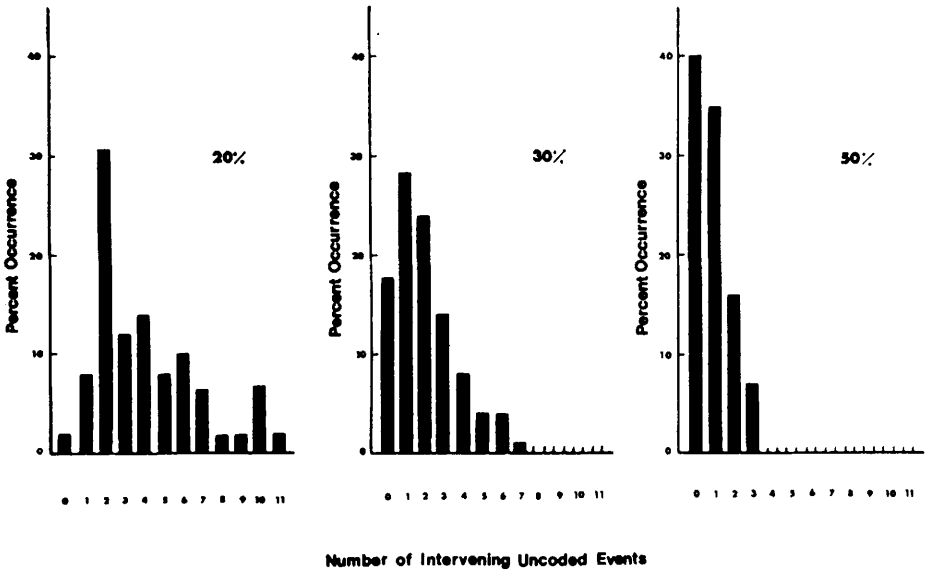

FIGURE 2. Number of events intervening between coded events for observers who coded 20%, 30%, or 50% of events, when sampling sessions on-line.

Interval Sampling

Procedures. Next, we evaluated the fidelity of frequency recording sampling procedures in which all events which occur within given time intervals spaced throughout longer treatment sessions were recorded, using the CIAS. The master records for five 10-minute videotapes of aphasia treatment sessions were sampled according to the following schedules: 6 minutes off (not sampled)/2 minutes on (sampled), 4 minutes off/1 minute on, 3 minutes off/2 minutes on, 2 minutes off/2 minutes on, 1 minute off/1 minute on. Each record sampled with a schedule was also sampled with the inverse of that schedule, for example, a record sampled at 4 minutes off/1 minute on, was also sampled at 1 minute on/4 minutes off, from the beginning of the master record for each tape. Because there were essentially no differences between a sampling schedule and its inverse (mean difference = 0.63%, SD = 0.14%), the records for the original and the inverted versions of each schedule were combined. Category-by-category absolute differences between the master record and each sampling schedule were computed, as before.

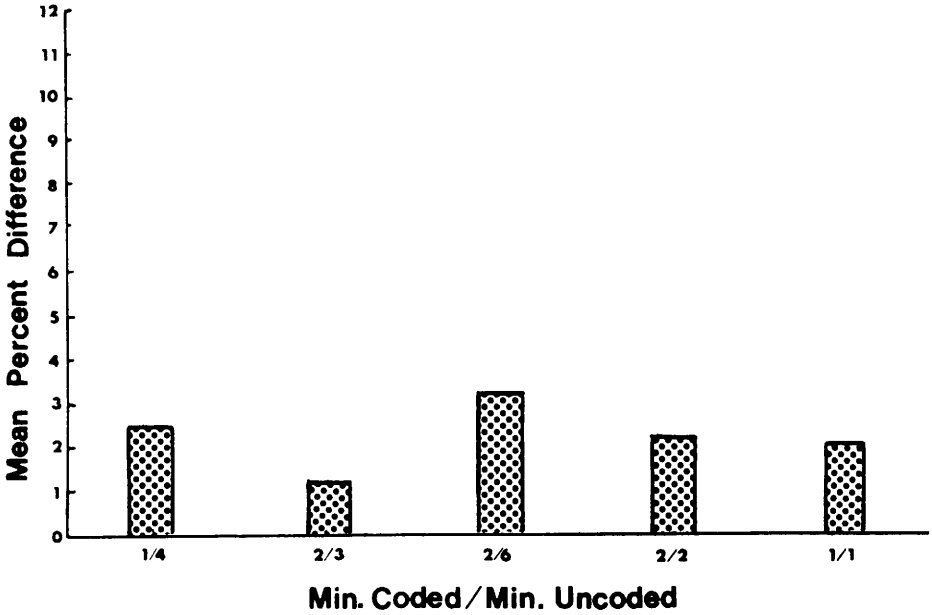

FIGURE 3. Mean percent difference from true proportions of events within treatment sessions for various multiple interval sampling ratios.

Results. The results of these computations are presented in Figure 3. Figure 3 shows that, even when only one minute of every four was recorded, records generated by recording during spaced intervals were highly representative of the records that were generated by exhaustively recording the whole treatment session. The gain in accuracy obtained by recording every other minute, rather than every fourth minute, was less than one percent.

The mean absolute differences from the master record and the standard deviations of those differences for each of the five tapes at 1/4, 1/1, and 2/6-minute sampling schedules are presented in Table 2. Once again standard deviations were generally small, so that the mean errors contained in Table 2 are representative of the 195 difference scores upon which the means were based. Ninety percent confidence intervals were calculated for each of the interval sampling procedures. The limits of these confidence intervals are presented in Table 2. Results indicated that when one samples using a 4 minute off/1 minute on schedule (the leanest schedule), all event categories sampled will be within 10% (absolute) of the true proportions of event categories which actually occur throughout the session ($p < 0.10$).

In order to evaluate the representativeness of single samples several minutes in duration taken from longer treatment sessions, we compared the records generated by sampling 2 minute or 5 minute segments of sessions taken either from the beginning or middle of a session with records generated by recording the entire session. The proportions of each event category obtained by sampling either the first 2 minutes, the first 5 minutes, minutes 5 and 6, or minutes

TABLE 2. Means, standard deviations, and 90% confidence intervals of absolute differences between master record and interval samples from that record with 1/4, 2/6 and 1/1 (coded/uncoded) minute interval samples.

(Coded/Uncoded) Samples	*Tape #* 1	2	3	4	5	*90% CI*
1/4						
Mean (%)	2.09	2.03	2.72	2.10	2.73	0-9.58
SD	3.79	5.57	4.02	4.89	3.70	–
2/6						
Mean (%)	2.06	4.19	3.99	2.10	3.21	0-9.90
SD	3.08	5.62	5.59	2.81	3.49	–
1/1						
Mean (%)	1.57	0.78	4.37	1.38	2.57	0-7.79
SD	2.54	1.82	7.01	1.98	3.80	–

3 through 8 of 5 ten-minute videotaped samples of treatment were compared, by computing absolute differences between master records and sampling records, in the manner described previously. The results of these analyses are presented in Table 3.

Single 2-minute or 5-minute samples generated error ranging from approximately 2% to approximately 10% deviation from the true proportions of events throughout the 10-minute session, with longer sampling intervals generally resulting in greater accuracy. Ninety percent confidence intervals were calculated on the data for single-interval samples. The limits of these confidence intervals are presented in Table 3. These confidence intervals suggest that when one samples a single 2-minute interval from a ten-minute sample, 90% of event categories recorded will be within 21% (absolute) of true category proportions, and that when one samples a single 5-minute interval, 90% of event categories recorded will be within 12-14% (absolute) of true category proportions. The data in Table 3 suggest that if one samples 2-minute intervals from 10-minute treatment sessions, 81% of event categories sampled will meet the 10% error criterion described earlier. If one samples 5-minute intervals from those tapes, 90% (approximately) of event categories sampled will meet the criterion.

Since our single 2-minute or 5-minute samples were taken from sessions which were only ten minutes long, it seemed likely that the results obtained might not be valid for longer treatment sessions. Therefore, we gathered five videotapes of treatment which ranged in length from 35 to 50 minutes (mean = 43 minutes). Each videotape was coded in its entirety, using the CIAS, and a master record was prepared for each videotape, in the manner described earlier. We then compared the data obtained by coding only the first five minutes, the middle five minutes, or the last five minutes of each tape. The results of these sampling procedures are presented in Table 4.

It can be seen in Table 4 that single 5-minute samples from 35 minute to 50 minute treatment sessions generated larger confidence intervals than any other

TABLE 3. Means, standard deviations, and 90% confidence intervals of absolute differences between master record and interval samples from that record for 0-2 minutes, 0-5 minutes, 3-8 minutes sampled intervals.

	Tape #					
Interval	*1*	*2*	*3*	*4*	*5*	*90% CI*
Two Minute						
Min 0-2						
Mean (%)	3.69	9.65	5.13	6.71	9.96	0-21.70
SD	6.29	11.50	5.40	9.89	11.39	–
Min 5-7						
Mean (%)	6.97	4.27	8.20	9.26	7.22	0-21.89
SD	7.49	5.14	14.54	9.28	8.14	–
Five Minute						
Min 0-5						
Mean (%)	2.65	3.07	3.97	5.20	7.54	0-13.68
SD	2.96	3.81	4.19	7.24	9.66	–
Min 3-8						
Mean (%)	2.95	4.55	6.22	2.26	3.73	0-11.74
SD	3.42	5.48	9.21	2.14	3.41	–

sampling method. Even the best 5-minute sampling procedure (middle 5 minutes) generated high predicted sampling error; the upper limits of 90% confidence intervals for event-by-event percent error extended from 24 to 33%. Only 85% (approximately) of event categories sampled would meet the 10% error criterion established for all sampling schedules.

In order to compare the relative accuracy of the three sampling procedures evaluated in this investigation, the 90% confidence intervals for each sampling procedure (momentary sampling at 10%, 20%, and 50% schedules; multiple interval sampling at 1/1, 2/6, 1/4 minute schedules; and single interval sampling of two-minute or 5-minute intervals) were graphed (Figure 4). It can be seen

TABLE 4. Means, standard deviations, and 90% confidence intervals for differences between master record and single-interval 5-minute samples from beginning, middle, and end of 35-50 minute sessions.

	Tape #					
Sampled Interval	*1*	*2*	*3*	*4*	*5*	*90% CI*
First 5 minutes						
Mean (%)	11.05	7.23	7.46	6.80	8.32	0-27.67
SD	13.92	11.84	11.81	9.37	12.16	–
Middle 5 minutes						
Mean (%)	6.35	4.47	8.82	7.50	6.54	–
SD	11.62	6.72	14.64	9.94	9.26	0-23.97
Last 5 minutes						
Mean (%)	10.28	10.10	8.89	8.70	8.49	–
SD	12.79	15.21	16.07	12.70	13.85	0-32.59

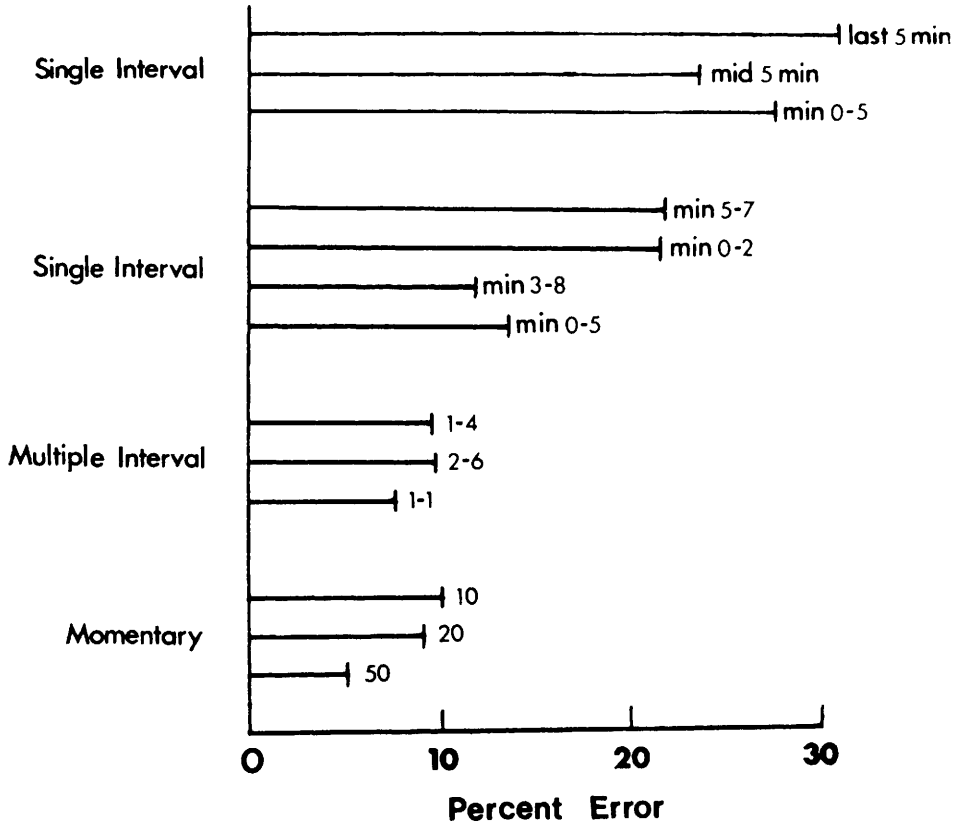

FIGURE 4. Confidence intervals for momentary, multiple interval, and single interval sampling schedules evaluated in this study.

in Figure 4 that predicted error limits for momentary and multiple-interval sampling are consistently smaller than the error limits for single-interval sampling. None of the momentary or multiple–interval sampling procedures generated predicted error ranges greater than 10% on either side of zero, while all four single-interval procedures generated greater than 10% positive error range.

These results mean that Olsen's (1972) and Schubert and Laird's (1975) suggestions that single 5-minute or 3-minute samples of treatment sessions will accurately represent the content of a complete session must be viewed with caution. Even though mean error in prediction using a single interval estimate of treatment content consistently averaged less than 10% per 10-minute tape, confidence intervals for single interval estimates were more than twice as wide as the confidence intervals for multiple-interval or momentary sampling methods. When a single 5-minute sample was taken from 30-minute and longer treatment sessions, the representativeness of the samples obtained was even poorer, with 15% of all event categories sampled exceeding 10% absolute error from true event proportions.

In general, then, the data indicate that periodic momentary sampling and multiple-interval sampling generate records with less error than single-interval sampling methods. If one were to use a single-interval sampling method, then one should attempt to sample as much of the session as possible during the sampled interval, because longer intervals will generate less sampling error than shorter ones.

CONCLUSIONS

Our analysis of several procedures for using samples taken from speech and language treatment sessions to represent the content of an entire session leads to several conclusions.

(1) A single 3–min or 5–min sample, taken from a longer (30–min or more) treatment session, will not, in a large proportion of cases, generate accurate representations of the longer session.

(2) A series of 1–min duration samples, taken periodically throughout the session, generates records which are highly accurate representations of the entire session. Less than 3% error is generated by sampling only every fifth minute.

(3) Sampling procedures which result in the recording of every n^{th} event generate records which are highly accurate representations of treatment sessions. Less than 10% error is generated by sampling, on the average, only every fifth event. If observers can reliably identify and record exactly every n^{th} event, then sampling only every tenth event will generate less than 3% error.

(4) Trained observers can be expected to record from 20% to 85% of events which occur in treatment sessions when they use coding systems such as the 39–category Clinical Interaction Analysis System to code those treatment sessions on–line and without recourse to videotape recordings, audio recordings, or transcripts.

(5) The primary requirement for accuracy of any sampling procedure is that the sampling procedure sample events at intervals distributed throughout the treatment sessions. Any method which restricts sampling to a restricted portion of the treatment session is susceptible to potentially unacceptable degrees of error in representing the treatment session. Even when events are sampled throughout a session, it might be possible to obtain a nonrepresentative sample, if events within the session were to occur according to regular temporal cycle and the sampling interval happened to match that cycle. If the sampling intervals were to vary around some mean value, the probability of event cycles and sampling intervals adventitiously matching would be extremely low.

ACKNOWLEDGMENT

This research was supported by National Institutes of Neurological and Communication Disorders and Stroke Contract 74-09, and by the Research Service, Minneapolis Veterans

Administration Hospital. Requests for reprints should be directed to R. H. Brookshire, Director, Aphasia Section (127A), Veterans Administration Hospital, 54th Street and 48th Avenue South, Minneapolis, Minnesota 55417.

REFERENCES

BALES, R. F., A set of categories for the analysis of small group interaction. *Amer. Sociol. Rev.*, **15**, 257-263 (1950).

BALES, R. F., *Personality and Interpersonal Behavior.* New York: Holt, Rinehart, and Winston (1970).

BIJOU, S. W., PETERSON, R. F., and AULT, M. H., A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *J. appl. Behav. Analysis*, **1**, 175-191 (1968).

BOONE, D. R., and PRESCOTT, T. E., Content and sequence analyses of speech and hearing therapy. *Asha*, **14**, 58-62 (1972).

BOONE, D. R., and PRESCOTT, T. E., *Speech and Hearing Therapy Scoring Manual: A Manual for Learning to Self-Score the Events of Therapy.* Grant No. OEG-0-70-4758-607, U.S. Department of Health, Education, and Welfare, Division of Research, Bureau of Education for the Handicapped, Office of Education (1974).

BROOKSHIRE, R., A system for coding and recording events in patient-clinician interaction during aphasia treatment sessions. In R. H. Brookshire (Ed.), *Clinical Aphasiology: Conference Proceedings, 1976.* Minneapolis, Minn.: BRK Publishers (1976).

CROMWELL, L., and SCHEIDEL, T. M., Categories for analysis of idea development in discussion groups. *J. social Psychol.*, **54**, 155-168 (1961).

DIEDRICH, W. M., *Charting Speech Behaviors.* Lawrence, Kan.: University of Kansas (1973).

DIEDRICH, W. M., Training speech clinicians in the recording and analysis of articulatory behavior. Final Report, Grant OEG-0-071-1689, U.S. Office of Education (1976).

FINE, H. J., and ZIMET, C. N., A quantitative method of scaling communication and the interaction process. *J. clin. Psychol.*, **12**, 268-271 (1956).

FLANDERS, N. A., *Interaction Analysis in the Classroom.* Minneapolis, Minn.: University of Minnesota College of Education (1960).

FLANDERS, N. A., *Helping Teachers Change Their Behavior.* Final Report, NDEA Projects 172102 and 7-32-0560, U.S. Office of Education (1963).

FLANDERS, N. A., *Analyzing Teaching Behavior.* Reading, Mass.: Addison-Wesley (1970).

JOHNSON, T., The development of a multidimensional scoring system for observing the clinical process in speech pathology. Doctoral dissertation, University of Kansas (1969).

KUNZE, L., A program for training in behavioral observation. In A. Miner (Ed.), A symposium: Improving supervision of clinical practicum. *Asha*, **9**, 473-476 (1967).

MOWRER, D., BAKER, R., and OWEN, C., Verbal content analysis of speech therapy sessions. Paper presented at the Annual Convention of the American Speech and Hearing Association, Denver (1968).

POWELL, J., MARTINDALE, A., and KULP, S., An evaluation of time-sampling measures of behavior. *J. appl. Behav. Analysis*, **8**, 463-469 (1975).

POWELL, J., MARTINDALE, B., KULP, S., MARTINDALE, A., and BAUMAN, R., Taking a closer look: Time-sampling and measurement error. *J. appl. Behav. Analysis*, **10**, 325-332 (1977).

REPP, A. C., DIETZ, D., BOLES, S. M., DIETZ, S. M., and REPP, C. F., Differences among common methods for calculating inter-observer agreement. *J. appl. Behav. Analysis*, **9**, 109-113 (1976).

SCHUBERT, G. W., MINER, A. L., and TILL, J. A., *The Analysis of Behavior of Clinicians System.* Grand Forks, N.D.: University of North Dakota (1973).

SCHUBERT, G. W., and LAIRD, B. A., The length of time necessary to obtain a representative sample of clinician-client interaction. *National Student Speech Hearing Assoc.*, **3**, 26-33 (1975).

THOMSON, C., HOLMBERG, M., and BAER, D. M., A brief report on a comparison of time-sampling methods. *J. appl. Behav. Analysis*, **7**, 623-626 (1974).

Received December 27, 1977.
Accepted July 21, 1978.

APPENDIX

Categories contained within the Clinical Interaction Analysis System

Dimension	*Categories*	
Type of Event	Imperative	Nonverbal
	Model	Explanation
	Completion	Clinician Discourse
	Yes-No Question	Patient Discourse
	Other Question	
Complexity of Request	Inference	
	Number of Words	
Manner	Spoken	
	Gestural	
	Melodic	
Materials	Object-Picture	
	Written	
Expected Response	Long	Written
	Spoken	Repeated
	Melodic	Delayed
	Gestural	
Support	Spoken Unison	
	Gestural Unison	
Patient's Response	Response	
	Requests Information	
	Normal Response	
	Unacceptable Response	
Clinician Feedback	Feedback	Negative Feedback
	Spoken Feedback	Correction Feedback
	Gestural Feedback	Repetition Feedback
	No. of Words in Feedback	Elaboration Feedback
	Positive Feedback	Intense Feedback