RELIABILITY OF OBSERVATION AS A FUNCTION OF RECORDING PROCEDURE, NUMBER OF CATEGORIES, AND LENGTH OF TIME INTERVAL: A METHODOLOGICAL STUDY

By

JO ANN BUDZIK

Bachelor of Science

Oklahoma Agricultural and Mechanical College

Stillwater, Oklahoma

1955

Submitted to the faculty of the Graduate School of
the Oklahoma Agricultural and Mechanical College
in partial fulfillment of the requirements
for the degree of
MASTER OF SCIENCE
May, 1956

RELIABILITY OF OBSERVATION AS A FUNCTION OF RECORDING PROCEDURE, NUMBER OF CATEGORIES, AND LENGTH OF TIME INTERVAL: A METHODOLOGICAL STUDY

Thesis Approved:

Henry D. Schalock
Thesis Adviser

Virginia Messenger Stapley

Josephine Hoffer

James Thacker

Robert MacVicar
Dean of the Graduate School

ACKNOWLEDGMENTS

The author wishes to express her indebtedness and sincere appreciation to Assistant Professor Henry D. Schalock at whose suggestion and under whose direction this study was made; to Mrs. Josephine Hoffer, Assistant Professor, and Dr. James Walters, Associate Professor, for their careful and critical reading of the manuscript; to Evelyn Miller, a graduate student, for her aid and cooperation as a trained observer in establishing reliability; and to the teaching staff of the nursery school-kindergarten program in the Department of Family Relations and Child Development at Oklahoma Agricultural and Mechanical College for their interest and cooperative attitude during the observations.

The author is also indebted to her husband, Jerry, for his encouragement during the time of the writing of this thesis.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Background of the Study

The use of the human observer as a measuring instrument in the behavioral sciences is becoming increasingly widespread. According to Heyns and Lippitt (8), the human observer as a measuring instrument has been most useful and necessary when other techniques of measurement would disrupt the process of social interaction and when the actor himself is inadequate as a direct source of information. In other words, the human observer has been relied upon when other devices for measurement are unavailable or inappropriate to the phenomena under investigation. While there are undoubtedly some unique problems in using the human observer as a measuring instrument, the objective is the same as that with any other measuring instrument, namely, to obtain accurate and reliable information.

The human observer has been used primarily to describe behavior as it occurs in response to observable external stimuli or situations. This research approach is based on the premise that behavior itself is a matter of immediate interest or value. Perhaps the greatest asset of this approach is that it permits the recording of behavior simultaneously with its spontaneous occurrence, and does not depend on retrospective or anticipatory reports (15).

However, together with these assets that the human measuring instrument offers, one persistent problem that reoccurs is that of reliability.

Reliability is defined as the degree to which two independent observers observing simultaneously can agree on the occurrence of the same events in time. Heyns and Lippitt point out that observer load is a factor affecting reliability of observation. They say that, "Other things being equal, the more the observer has to do, the lower the reliability" (8, p 397). In her review article on time sampling, Arrington points out that the most important factors affecting accuracy of observation in uncontrolled life situations are the amount of behavior observed, the degree of precision with which the observed behavior is defined, and the simplicity or complexity of the method of recording. She summarizes her point of view by saying:

> Other things being equal, the fewer the behavior items or categories included in the record, the more precise the definition of these items, and the simpler the recording process, the more reliable will be the observations (2, p 92).

Evidence from studies employing content analysis also suggest that reliability scores decrease as the number of categories used in the analysis are increased (10). There is some evidence, too, that indicates that the less inference required of observers in categorizing the behavior under observation the higher the degree of agreement (8). The degree of reliability attained may also be a function of the amount of training that the observers have had. If the observer is thoroughly trained, many of the sources of unreliability will have been dealt with in the training process.

In view of these factors affecting the reliability of observation, certain steps have been taken to enhance the accuracy of observational records. The observer's task has been simplified by limiting the number of kinds of behavior to be observed and the number of behavioral items or categories to be included in any one record, by defining the

behavioral items or categories to be used in precise, objective terms, and by substituting prepared record blanks and code symbols for longhand descriptions. In an attempt to increase the comparability of the observational data, observations have been made in terms of time intervals. This has provided a basis for an objective comparison of behavioral frequencies. However, it introduces another variable that affects the reliability of observation, namely, that of having to place a given behavior in the time interval in which it occurred. The reliability problem that this procedure introduces is that the smaller the time interval, and therefore the greater number of time intervals in any one period of observation, the greater the opportunity for observer error to result from disagreement between observers on the placement of a behavior in its appropriate time interval.

Thus we see that in spite of repeated efforts to simplify the observer's task, it still is a difficult one since it involves an instantaneous classification of each item of behavior at the time it is observed, precise timing of each occurrence of relevant behavior, and an accurate recording of the appropriate symbol in the appropriate place on the record blank (3).

This discussion suggests that the number of categories and the length of the time interval are factors which affect the reliability of observation. One other factor that may play an important role in reliability measurement is that of the method of recording observations. [This has OMIT been suggested by the experience of a group of students who were enrolled in a course in observational methods at Oklahoma A. & M. College in the fall of 1955. In this course students had practice in recording category symbols by means of a "check sheet record" recording method and a "running

OMIT

record" recording method. The consensus was that the attention required of the observer to locate the category placement on the check sheet record could cause a lowering of reliability. This was felt especially to be the case when the number of categories used in the observation was increased.

STOP

Purpose of the Study

The purpose of this study was to investigate the effect of three factors, i.e., number of categories, length of time interval, and method of recording observations, on reliability measurement. In this respect it was hypothesized that as the number of categories used in the observation increased, the reliability of observation would decrease. This has been suggested by the fact that as the number of categories are increased the load of the observer increases, i.e., as the number of categories increase the number of discriminations required of the observer to immediately classify the behavior into one of the categories used in the observation increases. It was also hypothesized that as the time interval in which the behavior was to be recorded decreased the reliability of observation would decrease. This has been suggested by the fact that as the length of the time interval decreases more and more observer attention must be focused upon making time discriminations, and by the fact that greater opportunity for "timing errors," i.e., when two observers that are observing simultaneously place the same incident of behavior in different time intervals, will appear with short time intervals. A further hypothesis was that the method of recording would affect reliability measurement, reliability decreasing faster with the check sheet record than it would with the running record as the number of categories were increased in the observation.

In summary, the purposes of this investigation were three:

1. To determine if reliability of observation varied with the method of recording when the number of categories and the length of the time interval used in the recording were held constant.

2. To determine if reliability of observation decreased with a decrease in the length of the time interval when the number of categories and the method of recording were held constant.

3. To determine if reliability of observation decreased with an increase in the number of categories used in the observation when the length of the time interval and method of recording were held constant.

The hypothesis guiding this study was that as the number of categories increased and the length of the time interval decreased, reliability of observation would decrease with both a running record and a check sheet record, but would decrease more with a check sheet record.

Review of Related Literature

The literature pertaining directly to the study presented herein was reviewed in the section headed Background of the Study. The literature that will be reviewed in this section pertains to the methodology that is frequently associated with the use of the human observer as a measuring instrument, namely, time sampling methodology.

Controlled observation or time sampling is a method of observing the behavior of individuals or groups under ordinary conditions of everyday life in which observations are made in a series of short time periods so distributed as to afford a representative sampling of the behavior under observation. It is a method the essential function of which is accurate measurement of the incidence of specified behavioral acts or patterns under specified conditions (2).

Time sampling owes its development to a series of related studies made at the University of Minnesota and to a sociological research program at the Columbia Child Development Institute. Olson, in 1926-27 (11), introduced the method of time sampling when he developed a technique for measuring nervous habits in populations of public school children. He felt that to attempt to record every incident of a given sort of behavior for a long period of time was an almost impossible task, and that it would be much easier to set up an arbitrary time interval for observation and to check whether or not some specific pre-defined behavior occurred in that time interval. Anderson, in the foreword to Olson's monograph on the measurement of nervous habits, characterized this research approach as:

> a method of time sampling, the essence of which lies in the observation of the behavior of each individual in respect to the particular category of activity upon which information is to be obtained, during a period of time that is kept constant for each individual observed (11, Foreword).

Olson's study of nervous habits demonstrated that such aspects of behavior could be defined and observed under controlled conditions. During the same year, Parten (12, 13, 14), used a modified form of the method devised by Olson to study social participation, leadership, and social play. Instead of using a five-minute time unit, Parten contributed to the discrimination of individual differences by using a one-minute time unit for measurement of frequency of behavior. Goodenough (6), further increased the discriminative value of the method by dividing the minute sample into fifteen-second intervals and by differentiating degrees of manifestation of each behavioral trait. In summarizing the work with this technique, Goodenough (7) has stated that the method of time sampling lends itself to all ordinary forms of statistical treatment, may be used

by persons with only a moderate degree of training, is not excessive in its time requirements, and may be adapted to the study of many different forms of behavior. The method consists simply of the observation of the everyday behavior of an individual or group of individuals for definite short periods of time and the recording of the occurrence or non-occurrence of certain specified and objectively defined forms of behavior during each of these periods.

Meanwhile, at Columbia, Thomas became interested with the apparent need for more reliable and comparable measures of behavior frequency than could be derived from traditional diary records. A series of studies under Thomas' direction (1, 4, 5, 9, 16, 17), were undertaken to investigate some of the specific problems involved in observer reliability. One of these studies (9), provided a unique contribution in sampling methods in that an attempt was made to apply the criteria of random sampling to observation of social behavior.

From this brief review of time-sampling methods, it can be seen that the technique of time sampling has been applied in a variety of ways in studies of child behavior. Some investigators have observed carefully pre-defined aspects of social behavior while others have recorded all behavior without previous definition. Some have recorded their observations in code, others in longhand. Some have been concerned with the frequency of certain behavioral acts, and others, with the classification of these acts into categories. Some studies have been focused on the individual, others on a group of individuals. Length of the time sample has varied depending on the type of behavior sampled and the purpose of the sampling. Some studies have measured the occurrence or non-occurrence of the behavior under observation within the time sample, while others have measured the frequency of occurrence within

the sample (2). The range and usefulness of techniques of this type are difficult to determine, however, for they have been applied to small groups and results have been affected by too many variable factors.

It can be said with certainty, however, that the time sampling method does offer distinct advantages over some of the other approaches to observation. It is an objective method which can easily be applied to studies of normal behavior in ordinary life situations; it permits a carefully selected and reported sample of observations, and it permits quantitive results capable of statistical analysis. Possibly, the only disadvantage is that in order to secure satisfactory measures a considerable expenditure of time is required.

CHAPTER II

PROCEDURE

The methodology used in this study was that of direct observation. The reliability of observation was determined by two observers who worked together throughout the study. The details of the procedure used in the study are outlined below.

One child was observed in this study. By using one child, one important source of variation affecting reliability measures was controlled, namely, the characteristic differences in speed or rate or kind of behavior that appear in different individuals. If more than one child had been observed, it would not have been possible to determine whether or not variations in observer reliability scores were a function of the controlled conditions of observation or of variations in the kind or rate or speed of behavior the different children observed had evidenced.

The child observed in the study was selected from the three-year-old children enrolled in the Oklahoma Agricultural and Mechanical College Nursery School-Kindergarten Laboratory. To insure that the child selected for observation would evidence the kinds of behavior under observation, the three-year-old group teacher and her student teachers were presented the categories that were to be used in the observations and asked to place in rank order the three children in the group who would be most likely to evidence these kinds of behaviors. The child ranked first by the majority of teachers was chosen as the subject for the study. When

the observations for the study were half finished, the subject selected for the study dropped from the group and entered another group. The child ranked second by the teachers was then observed for the remaining half of the study.

Since this study was a methodological one and involved no generalizations to a population, the sampling procedure focused upon obtaining the kinds of behavior that appeared frequently enough to allow meaningful tests of reliability, rather than upon population representativeness.

Method of Observation

The child was observed as he interacted with the children and adults in the nursery school group. The observations were made in terms of categories. The categories used consisted of the twenty categories of interactive behavior (one of these categories referred to behavior that was non-interactive) that appeared most frequently in Schalock's study (15). In some instances, these categories have been slightly modified to better suit this study. These particular categories were selected because of the premium this study placed upon the frequency of category appearance.

The categories, their symbols and definitions, appear below. Scoring symbols appear to the left of each category. In the category description, A=Adult and C=Child.

o Non-Attendance: C directs his attention to something other than A or C. Ex: C plays quietly with back to A or C, or C stares out the window. Non-Attendance may be accompanied by humming, singing, or talking, but this behavior is not apparently directed toward A or C. No interactive characteristics are observable.

e Recognition: C responds to A's or C's stimulation in a way that

indicates definite awareness of A's or C's statement or activity. Ex: Mm hm. Ex: Yes, I see.

r Restricting: C attempts to modify A's or another C's behavior by reducing the intensity, or speed, or manner of executing the behavior, etc., but apparently does not intend to stop the activity completely. Ex: Be careful or you will fall and hurt yourself. Ex: I don't think you should pound the table quite so hard.

d_2 Directing by a Command: C attempts to influence A or another C by giving a command. Ex: Get me that book.

e→ Seeking Recognition: Attention to activities, productions or statements is sought by C. Ex: See what I've made.

c Attendant Observation: C noticeably directs his attention to A or C and/or A's or C's activity by silently watching. Ex: C watches another C draw a picture.

p Joint Participation in Activity: A and C or C and C are mutually engaged in the same activity. Ex: C and C are reading, working with puzzles, singing, playing, drawing, or painting together.

f Forbidding: An ongoing activity is interfered with by C with the apparent intent of stopping it completely. Ex: Stop putting my toys away. I want to play with them. Ex: Stop that. Ex: You can't have that chair, but you can sit on the other one.

d_1 Directing by Suggestion: C attempts to influence A or another C by using verbal or nonverbal suggestion. Ex: You could play with the darts if you wanted to. Ex: Would you like to sit over here?

h→ Seeking Help: C requests physical assistance. This may be done either verbally or non-verbally. Ex: Help me carry these chairs. Ex: I can't pound this right. Will you hold this peg for me?

b Statement of Condition or Action: C comments describing an existent state, situation, or an action. Ex: It's hot in here. Ex: I'm afraid. These statements are a type of information but do not appear relevant to the solution of problems or making decisions. Though statements of condition or action may have stimulus properties, from the observer's point of view, they do not apparently influence behavior or elicit particular responses.

t Offering Information: C offers knowledge or guidance verbally or by means of demonstration. Ex: You have to unsnap my boots before you can take them off. Ex: C shows another C how to make wheels with tinker toys when he is having difficulty.

+1 Cooperation: C responds to A's or C's comments, suggestions, or requests with apparent interest and willingness. Ex: I'll play darts with you. Ex: I would love to play house with you. What could I do?

w_1 Rejection by Changing Subject: C changes the subject or interrupts A or C with irrelevant conversation. Ex: A: Pick up the blocks. C: It's cold in here, isn't it?

q→ Seeking Permission: Consent is sought by C for a proposed activity. Ex: Is it all right if I leave for a minute?

t→ Seeking Information: C asks for information. This information may be personal or information sought concerning an ongoing activity. Ex: Does your little brother break your toys? Ex: Do you like that puppet more than the others?

w_4 Rejection by Ignoring or Evading: A or C makes a direct stimulation to C and C responds by giving some indication of ignoring or evading. Ex: C looks at C and says, "Do you like me?" C does not respond. Ex: C: Don't you think this is a nice picture I have made for you? Other C does not respond.

-1 Non-Cooperation: Refusal to accept commands, suggestions, or requests. Ex: I won't play in the water with you, but I'll hand you the boats. Ex: No!

w_3 Rejection as a Person: C rejects A or C as a person. Ex: I don't want anything to do with you. Get away from me.

w_2 Rejection by Denying the Validity of Statement: C denies the validity of A's or C's statement or action. Ex: C: I'm a good boy. C: No, you are a bad boy. Ex: C shakes head negatively.

Conditions of Observation

Observations were made and tested for reliability under each of the following conditions.

1. Running Record: In this study a "running record" referred to the record resulting from the continuous recording of the symbols representing categories of behavior on a prepared record blank. The categories were committed to memory and as behavior occurred which could be identified by a category it was recorded on the record sheet.

Two factors were varied systematically in the observations using the running records, the number of categories and the length of the time interval. A time interval referred to the interval of time in which observations were recorded. This was distinct from the length of the time unit of behavior. For example, if a one-minute time interval were

used and the length of the time unit was fifteen minutes, there would be fifteen time intervals within the time unit.[1]

Using five categories, observations were made for time intervals of one minute, thirty seconds, fifteen seconds, and five seconds. Two fifteen-minute periods of observation were used to demonstrate reliability for each of these time intervals.

A similar procedure was used for observations involving ten, fifteen, and twenty categories. To increase the number of categories the observers were using, groups of five categories were added to the categories used in the previous series of observations. An important methodological consideration was raised at this point. If the hypothesis that reliability would decrease as the number of categories in the observation schedule was increased was to be tested, each set of five categories used in the observations had to contribute equally to the observer's load. If each of the groups of categories added to the observation schedule did not increase the load of the observer in an equivalent way it would not have been meaningful to speak in terms of five categories or ten categories or fifteen categories. For example, if five categories were added to the observer's load, yet only one of these five categories appeared in the subsequent observations, the addition of these five categories would not have increased the load of the observer in a way equivalent to the addition of five categories that appeared frequently in subsequent observations.

---

[1] Stopwatches were used throughout the study in keeping time. Each observer held a stopwatch in his hand for all of the observations. The observers began their observations simultaneously, and periodically throughout the observations they would call time to one another as a means of checking to see that they were together in time and that they were scoring in the same interval on the record blank.

In an attempt to make each set of five categories contribute equally to the observer's load, the twenty categories selected from Schalock's study (15) were divided into groups on the basis of their frequency of appearance. Data on the frequency of appearance of each of these categories were available. Observing these data, each of the four groups of categories were made up of categories that appeared with approximately the same frequency. The four most frequently appearing categories were divided so that each of the groups would receive one of these categories. The four categories appearing next most frequently were similarly apportioned, and so on.

2. Check Sheet Record: In this study a "check sheet record" referred to the record resulting from the continuous checking of categories of behavior as these categories appeared on a prepared record blank. The categories to be checked were listed on the record blank, and as behavior occurred that could be identified by a category, it was recorded by placing a check after that given category in the appropriate time interval.

The number of categories and the length of the time interval used were varied for the check sheet record observations in the same way as they were varied for the observations with the running record.

Reliability of Observation

Reliability data were obtained for each of the above conditions of observation in order to determine the effect each of these conditions had on the demonstration of observational reliability. These reliability data are presented in terms of the percentages of agreement between two persons observing simultaneously but independently. The procedure used by Schalock (15) in computing category and observer reliability was used in this study. The formula was:

$$\frac{\text{number of agreements}}{\text{number of agreements plus number of disagreements}}.$$

Procedure of Observation

Two observers took five categories and committed them to memory. They began practice observations with these categories using the running record procedure with a time interval of one minute.

Observations began with the running record rather than with the check sheet record in order that a known level of observer familiarity with the categories was assured, i.e., the level of familiarity required to establish 80 per cent agreement between two observers on the identification of each category, and therefore on the over-all reliability of observers. Given this level of familiarity with the categories for the running record it was assured that the observers were at least this familiar with the categories when using the check sheet record. Looking at it in another way, if the observers had started their observations with the check sheet record, it would have been possible that they would have been able to demonstrate adequate reliability for these observations, and yet not have been able to demonstrate reliability of observation on the running record without additional practice. It may be seen, then, that starting the observations with the running record was simply a means of insuring that the observers were equally capable of reliable observation with the check sheet record and the running record.

Practice observations with the running record using five categories and a time interval of one minute were undertaken and continued until the two observers could demonstrate at least 80 per cent agreement on each

of the categories used in the observation.[2] Again, this meant that an observer agreement of .80 was insured. Preliminary reliability data appear in Tables I and II.

After satisfactory reliability had been established for five categories using the running record and a one-minute time interval, the observations for purposes of data collection began. These observations began with the same conditions for which reliability was demonstrated, namely, the running record, five categories and a one-minute time interval. Both observers observed under these conditions for two fifteen-minute periods of time. After these observations had been completed, the time interval was changed to thirty seconds and the two observers again observed for two fifteen-minute periods. The observations using a thirty-second time interval were not preceded by practice observations since the purpose of these data were to determine the influence of decreased time intervals on reliability, given a known level of reliability at the outset. The same procedure was followed for the fifteen-second and five-second time intervals. The data for the study consist of the observer reliability scores obtained for each of the above observations.

When the above observations were completed, the same five categories were used in the observations with the check sheet record. Practice observations again using a one-minute time interval were undertaken. When adequate category reliability had been established the observations

---

[2]After several practice observations using five categories and the check sheet record, the observers were able to demonstrate only .75 agreement on one of the categories. Although this figure was below the arbitrary reliability score of .80, it was decided to proceed with the observations for purposes of data collection on the assumption that with continued practice an observer agreement of .80 on this particular category would be reached.

TABLE I

RELIABILITY DATA FOR INDIVIDUAL CATEGORIES
USING THE RUNNING RECORD

Category	Frequency	Per cent of Agreement
Non-Attendance	10	90
Recognition	2	50
Restricting	3	100
Directing by a Command	7	100
Seeking Recognition	1	100
Total Reliability for Categories		96
Attendant Observation	7	86
Joint Participation in Activity	16	94
Forbidding	2	100
Directing by a Suggestion	8	86
Seeking Help	1	100
Total Reliability for Categories		91
Statement of Condition or Action	14	80
Offering Information	1	100
Cooperation	1	100
Rejection by Changing Subject	1	100
Seeking Permission	1	100
Total Reliability for Categories		83
Seeking Information	7	86
Rejection by Ignoring or Evading	3	100
Non-Cooperation	1	100
Rejection - As a Person	1	100
Rejection by Denying Validity of Statement	2	100
Total Reliability for Categories		93

TABLE II

RELIABILITY DATA FOR INDIVIDUAL CATEGORIES
USING THE CHECK SHEET RECORD

Category	Frequency	Per cent of Agreement
Non-Attendance	10	100
Recognition	2	100
Restricting	4	75
Directing by a Command	5	80
Seeking Recognition	1	100
Total Reliability for Categories		91
Attendant Observation	14	81
Joint Participation in Activity	16	82
Forbidding	2	100
Directing by a Suggestion	2	100
Seeking Help	1	100
Total Reliability for Categories		83
Statement of Condition or Action	17	82
Offering Information	8	88
Cooperation	2	100
Rejection by Changing Subject	2	100
Seeking Permission	1	100
Total Reliability for Categories		87
Seeking Information	9	89
Rejection by Ignoring or Evading	3	100
Non-Cooperation	2	100
Rejection - As a Person	1	100
Rejection by Denying Validity of Statement	1	100
Total Reliability for Categories		93

with the check sheet record for purposes of data collection began. These observations followed the same procedure as described for the running record, namely, both observers observing for two fifteen-minute periods using time intervals of one minute, thirty seconds, fifteen seconds, and five seconds.

When all of the conditions of observation using five categories were completed, five new categories were added to the original list. Before observations for purposes of data collection with the ten categories began, reliability was demonstrated for the new categories. This involved a procedure similar to the one described for beginning the observations with the five categories, namely, committing the new categories to memory and using the running record with a time interval of one minute. Practice was carried on with these categories until satisfactory reliability was demonstrated. When this level of proficiency was reached, the observers began their observations for purposes of data collection with the ten categories. The same procedure of observation as that described for five categories was then carried out with these ten categories. A similar procedure was followed using fifteen and twenty categories.

CHAPTER III

RESULTS

The three problems investigated in this study were (1) to determine if reliability of observation varied with the method of recording when the number of categories and the length of the time interval used in the recording were held constant, (2) to determine if reliability of observation decreased with a decrease in the length of the time interval when the number of categories and the method of recording were held constant, and (3) to determine if reliability of observation decreased with an increase in the number of categories when the length of the time interval and method of recording were held constant. The data relevant to these three problems appear in detail in Tables III and IV. It will be noted that these tables present the category frequency and percentage of observer agreement for each period of observation together with the total reliability score for both observations. The total observer reliability for each group of categories is also presented.

Data from these two tables have been summarized in reference to each of the three problems investigated in the study, and have been presented in conjunction with separate discussions of each of these problems.

Method of Recording

Table V contains the total reliability scores using the running record and the check sheet record with number of categories and length

TABLE III

FREQUENCY OF INDIVIDUAL CATEGORIES AND OBSERVER RELIABILITY SCORES OF OBSERVATIONS WITH THE RUNNING RECORD

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group I - 5 Categories										
Non-Attendance	11	100	11	100	100	24	96	32	93	95
Recognition	0	-	0	-	-	2	100	3	100	100
Restricting	1	100	1	100	100	0	-	0	-	-
Directing by a Command	10	90	5	100	94	0	-	3	100	100
Seeking Recognition	0	-	5	100	100	9	89	0	-	89
Total Observer Reliability		95		100	98		94		95	94
	Fifteen Seconds					Five Seconds				
Non-Attendance	23	96	28	87	90	119	82	142	91	87
Recognition	0	-	1	100	100	3	33	0	-	33
Restricting	2	100	0	-	100	1	100	0	-	100
Directing by a Command	4	75	3	100	86	4	25	2	50	33
Seeking Recognition	4	75	0	-	75	11	73	3	100	78
Total Observer Reliability		91		89	90		79		91	85

TABLE III (Continued)

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group II - 10 Categories										
Non-Attendance	12	83	24	86	86	28	83	21	84	84
Recognition	0	-	1	100	100	1	100	2	100	100
Restricting	3	100	1	0	75	0	-	1	100	100
Directing by a Command	2	100	4	100	100	0	-	5	100	100
Seeking Recognition	1	100	3	100	100	1	0	2	50	33
Attendant Observation	10	100	14	93	96	14	82	10	80	80
Joint Participation in Activity	2	50	4	50	50	15	89	20	90	89
Forbidding	1	100	0	-	100	1	100	1	100	100
Directing by a Suggestion	2	100	0	-	100	1	100	1	100	100
Seeking Help	0	-	0	-	-	4	100	2	100	100
Total Observer Reliability		91		88	89		85		88	86
	Fifteen Seconds					Five Seconds				
Non-Attendance	43	86	14	86	86	85	67	105	88	78
Recognition	1	100	3	100	100	1	100	4	75	80
Restricting	0	-	0	-	-	0	-	1	100	100
Directing by a Command	1	100	2	100	100	0	-	3	100	100
Seeking Recognition	3	100	0	-	100	4	100	1	0	80
Attendant Observation	30	80	10	60	75	98	70	35	60	67
Joint Participation in Activity	2	100	41	90	91	21	76	69	82	81
Forbidding	1	100	1	100	100	0	-	2	50	50
Directing by a Suggestion	8	100	10	100	100	3	100	4	75	86
Seeking Help	0	-	0	-	-	0	-	5	60	60
Total Observer Reliability		87		88	87		72		81	76

TABLE III (Continued)

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group III - 15 Categories										
Non-Attendance	54	88	25	74	84	14	93	26	85	88
Recognition	0	-	3	100	100	1	0	0	-	0
Restricting	0	-	0	-	-	0	-	0	-	-
Directing by a Command	1	100	2	100	100	0	-	1	100	100
Seeking Recognition	0	-	5	80	80	0	-	3	100	100
Attendant Observation	50	86	25	76	83	35	93	38	97	96
Joint Participation in Activity	0	-	0	-	-	0	-	0	-	-
Forbidding	0	-	1	100	100	0	-	1	100	100
Directing by a Suggestion	0	-	0	-	-	0	-	1	100	100
Seeking Help	0	-	0	-	-	0	-	0	-	-
Statement of Condition or Action	8	64	5	80	70	2	50	5	80	71
Offering Information	0	-	0	-	-	0	-	0	-	-
Cooperation	0	-	0	-	-	0	-	0	-	-
Rejection by Changing Subject	1	100	0	-	100	0	-	0	-	-
Seeking Permission	0	-	0	-	-	0	-	0	-	-
Total Observer Reliability		86		77	83		92		92	92

TABLE III (Continued)

Category	Time Interval									
	Fifteen Seconds					Five Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group III - 15 Categories										
Non-Attendance	40	92	51	69	84	92	90	122	95	93
Recognition	3	100	0	-	100	2	50	1	0	33
Restricting	0	-	0	-	-	0	-	0	-	-
Directing by a Command	0	-	1	100	100	6	100	0	-	100
Seeking Recognition	4	100	0	-	100	2	100	1	100	100
Attendant Observation	64	95	42	67	83	54	83	52	75	79
Joint Participation in Activity	0	-	1	100	100	14	90	4	75	88
Forbidding	0	-	0	-	-	19	90	0	-	90
Directing by a Suggestion	0	-	0	-	-	4	75	0	-	75
Seeking Help	0	-	0	-	-	0	-	1	100	100
Statement of Condition or Action	9	89	4	75	85	8	100	2	100	100
Offering Information	2	50	0	-	50	0	-	0	-	-
Cooperation	0	-	0	-	-	0	-	0	-	-
Rejection by Changing Subject	0	-	0	-	-	0	-	0	-	-
Seeking Permission	0	-	0	-	-	0	-	0	-	-
Total Observer Reliability		93		70	83		89		89	89

TABLE III (Continued)

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group IV - 20 Categories										
Non-Attendance	24	75	16	75	75	24	79	16	81	80
Recognition	6	67	3	100	78	3	67	3	100	83
Restricting	1	100	0	-	100	0	-	1	100	100
Directing by a Command	4	100	2	100	100	1	100	3	67	75
Seeking Recognition	2	50	1	100	75	1	100	2	100	100
Attendant Observation	38	87	22	86	87	35	80	29	66	73
Joint Participation in Activity	1	100	0	-	100	11	73	20	85	81
Forbidding	2	100	2	100	100	2	100	1	100	100
Directing by a Suggestion	3	100	1	100	100	1	100	2	100	100
Seeking Help	0	-	0	-	-	0	-	0	-	-
Statement of Condition or Action	11	82	6	67	77	15	80	12	67	74
Offering Information	1	100	0	-	100	2	100	0	-	100
Cooperation	0	-	1	100	100	1	100	0	-	100
Rejection by Changing Subject	0	-	0	-	-	0	-	0	-	-
Seeking Permission	0	-	0	-	-	1	100	0	-	100
Seeking Information	3	100	0	-	100	5	100	4	100	100
Rejection by Ignoring or Evading	0	-	1	100	100	0	-	2	100	100
Non-Cooperation	0	-	0	-	-	1	100	0	-	100
Rejection - As a Person	0	-	0	-	-	0	-	0	-	-
Rejection by Denying Validity of Statement	1	100	0	-	100	0	-	2	100	100
Total Observer Reliability		84		84	84		81		78	80

TABLE III (Concluded)

Category	Time Interval									
	Fifteen Seconds					Five Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group IV - 20 Categories										
Non-Attendance	26	61	31	68	65	16	56	66	60	60
Recognition	3	33	1	100	50	2	0	2	50	33
Restricting	0	-	1	100	100	1	100	0	-	100
Directing by a Command	20	85	13	77	82	1	100	0	-	100
Seeking Recognition	2	100	3	100	100	1	0	8	88	78
Attendant Observation	42	74	34	62	68	71	51	106	64	59
Joint Participation in Activity	22	81	3	67	80	112	58	16	62	59
Forbidding	5	100	3	67	88	0	-	0	-	-
Directing by a Suggestion	6	83	0	-	83	0	-	0	-	-
Seeking Help	2	100	0	-	100	0	-	0	-	-
Statement of Condition or Action	19	78	21	76	77	19	73	7	57	62
Offering Information	1	100	1	100	100	0	-	0	-	-
Cooperation	0	-	0	-	-	0	-	1	100	100
Rejection by Changing Subject	0	-	1	100	100	0	-	0	-	-
Seeking Permission	0	-	0	-	-	0	-	1	100	100
Seeking Information	1	100	7	56	63	3	67	1	0	50
Rejection by Ignoring or Evading	0	-	1	100	100	0	-	3	67	67
Non-Cooperation	0	-	1	100	100	0	-	0	-	-
Rejection - As a Person	0	-	1	100	100	0	-	0	-	-
Rejection by Denying Validity of Statement	0	-	6	100	100	0	-	0	-	-
Total Observer Reliability		77		72	74		56		64	60

TABLE IV

FREQUENCY OF INDIVIDUAL CATEGORIES AND OBSERVER RELIABILITY SCORES OF OBSERVATIONS WITH THE CHECK SHEET RECORD

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group I - 5 Categories										
Non-Attendance	14	93	16	84	87	22	79	17	90	82
Recognition	3	100	1	100	100	1	100	1	100	100
Restricting	1	100	1	100	100	1	100	1	100	100
Directing by a Command	4	100	4	75	88	0	-	2	100	100
Seeking Recognition	0	-	1	100	100	1	100	0	-	100
Total Observer Reliability		95		83	89		80		91	85
	Fifteen Seconds					Five Seconds				
Non-Attendance	48	90	32	86	89	84	86	19	95	87
Recognition	0	-	2	100	100	1	100	1	100	100
Restricting	1	100	0	-	100	0	-	0	-	-
Directing by a Command	0	-	2	100	100	1	100	2	50	67
Seeking Recognition	0	-	0	-	-	0	-	0	-	-
Total Observer Reliability		90		89	89		86		94	87

TABLE IV (Continued)

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both	Obs. 1		Obs. 2		Both
	Freq.	% of Agree.	Freq.	% of Agree.	Obs.	Freq.	% of Agree.	Freq.	% of Agree.	Obs.
Group II - 10 Categories										
Non-Attendance	12	83	25	100	97	26	84	42	95	90
Recognition	3	100	2	50	80	14	100	3	33	93
Restricting	0	-	0	-	-	0	-	0	-	-
Directing by a Command	2	100	1	0	50	1	0	1	0	0
Seeking Recognition	0	-	0	-	-	2	100	0	-	100
Attendant Observation	8	63	23	100	90	34	80	36	100	90
Joint Participation in Activity	13	100	2	100	100	13	70	8	88	76
Forbidding	0	-	9	77	77	2	50	0	-	50
Directing by a Suggestion	1	0	4	25	20	2	50	0	-	50
Seeking Help	1	100	1	0	50	1	100	0	-	100
Total Observer Reliability		88		88	88		81		93	87
	Fifteen Seconds					Five Seconds				
Non-Attendance	28	82	54	96	90	75	81	28	80	80
Recognition	4	75	6	100	90	11	55	4	100	67
Restricting	0	-	0	-	-	2	50	0	-	50
Directing by a Command	5	80	1	100	83	7	86	3	67	80
Seeking Recognition	2	50	7	72	65	0	-	0	-	-
Attendant Observation	34	71	42	69	70	77	90	33	80	86
Joint Participation in Activity	36	69	18	67	69	70	90	163	95	94
Forbidding	3	33	2	100	60	4	50	2	100	67
Directing by a Suggestion	10	60	3	67	61	1	100	0	-	100
Seeking Help	0	-	1	100	100	0	-	0	-	-
Total Observer Reliability		71		78	75		81		90	85

TABLE IV (Continued)

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group III - 15 Categories										
Non-Attendance	11	73	4	75	73	22	64	11	55	61
Recognition	2	100	4	75	83	7	57	11	73	67
Restricting	2	100	1	100	100	1	100	0	-	100
Directing by a Command	11	73	5	80	75	1	100	4	50	60
Seeking Recognition	2	100	1	100	100	1	100	3	100	100
Attendant Observation	17	82	16	81	82	33	76	37	65	70
Joint Participation in Activity.	1	100	3	100	100	0	-	5	80	80
Forbidding	4	75	1	100	80	2	100	3	67	80
Directing by a Suggestion	12	75	21	86	82	5	80	14	57	63
Seeking Help	1	100	1	100	100	4	50	0	-	50
Statement of Condition or Action	17	94	25	90	90	7	71	18	44	60
Offering Information	0	-	3	100	100	1	100	2	50	67
Cooperation	0	-	1	100	100	1	100	0	-	100
Rejection by Changing Subject	1	100	1	100	100	0	-	1	100	100
Seeking Permission	0	-	0	-	-	0	-	0	-	-
Total Observer Reliability		84		85	84		72		65	67

TABLE IV (Continued)

Category	Time Interval									
	Fifteen Seconds					Five Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group III - 15 Categories										
Non-Attendance	41	80	30	67	73	64	66	100	79	74
Recognition	12	75	4	75	75	8	63	2	50	60
Restricting	0	-	0	-	-	0	-	0	-	-
Directing by a Command	4	50	1	100	60	1	100	0	-	100
Seeking Recognition	5	80	0	-	80	3	67	3	67	67
Attendant Observation	43	72	60	80	77	141	80	81	68	76
Joint Participation in Activity	0	-	2	100	100	0	-	27	81	81
Forbidding	5	60	0	-	60	1	100	1	100	100
Directing by a Suggestion	1	100	1	100	100	9	44	0	-	44
Seeking Help	0	-	1	100	100	0	-	0	-	-
Statement of Condition or Action	7	57	15	60	59	25	68	11	73	70
Offering Information	0	-	0	-	-	1	100	1	0	50
Cooperation	2	50	0	-	50	0	-	0	-	-
Rejection by Changing Subject	2	100	0	-	100	1	100	0	-	100
Seeking Permission	1	100	1	100	100	0	-	0	-	-
Total Observer Reliability		73		74	74		73		75	74

TABLE IV (Continued)

Category	Time Interval									
	One Minute					Thirty Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group IV - 20 Categories										
Non-Attendance	29	90	14	93	91	38	74	18	83	75
Recognition	4	75	8	63	67	3	67	3	67	67
Restricting	0	-	1	100	100	1	100	1	100	100
Directing by a Command	5	100	15	67	75	1	100	3	100	100
Seeking Recognition	0	-	3	33	33	5	60	2	100	71
Attendant Observation	39	72	21	71	71	35	66	17	76	69
Joint Participation in Activity	0	-	12	75	75	2	100	18	72	75
Forbidding	1	100	2	50	67	1	100	1	100	100
Directing by a Suggestion	0	-	5	100	100	4	75	4	100	88
Seeking Help	1	0	1	100	50	1	100	3	67	75
Statement of Condition or Action	15	80	9	67	75	14	57	11	73	64
Offering Information	1	100	0	-	100	1	100	4	100	100
Cooperation	1	100	5	60	67	2	100	2	100	100
Rejection by Changing Subject	0	-	0	-	-	2	50	1	100	75
Seeking Permission	0	-	0	-	-	1	100	0	-	100
Seeking Information	4	100	2	50	83	3	67	0	-	67
Rejection by Ignoring or Evading	1	100	3	100	100	2	100	3	100	100
Non-Cooperation	0	-	1	100	100	0	-	3	100	100
Rejection - As a Person	1	100	0	-	100	0	-	1	100	100
Rejection by Denying Validity of Statement	1	100	0	-	100	0	-	1	100	100
Total Observer Reliability		82		74	79		71		82	77

TABLE IV (Concluded)

Category	Time Interval									
	Fifteen Seconds					Five Seconds				
	Obs. 1		Obs. 2		Both Obs.	Obs. 1		Obs. 2		Both Obs.
	Freq.	% of Agree.	Freq.	% of Agree.		Freq.	% of Agree.	Freq.	% of Agree.	
Group IV - 20 Categories										
Non-Attendance	18	78	27	63	69	28	64	26	50	57
Recognition	7	71	17	71	71	13	69	5	80	72
Restricting	3	67	0	-	67	0	-	0	-	-
Directing by a Command	1	100	0	-	100	11	73	2	50	69
Seeking Recognition	1	100	1	0	50	0	-	3	100	100
Attendant Observation	60	88	60	68	78	79	75	84	74	74
Joint Participation in Activity	0	-	0	-	-	42	76	61	82	79
Forbidding	0	-	0	-	-	1	100	3	100	100
Directing by a Suggestion	0	-	3	67	67	16	75	8	63	71
Seeking Help	1	100	0	-	100	1	100	0	-	100
Statement of Condition or Action	10	80	24	75	76	32	72	16	56	67
Offering Information	0	-	0	-	-	0	-	1	100	100
Cooperation	2	100	2	50	75	2	100	1	0	50
Rejection by Changing Subject	0	-	1	100	100	0	-	0	-	-
Seeking Permission	0	-	1	100	100	0	-	0	-	-
Seeking Information	7	56	6	50	54	12	83	2	100	86
Rejection by Ignoring or Evading	5	40	1	0	33	0	-	2	100	100
Non-Cooperation	1	100	3	67	75	0	-	1	100	100
Rejection - As a Person	0	-	1	100	100	0	-	2	50	50
Rejection by Denying Validity of Statement	1	0	0	-	0	1	100	3	100	100
Total Observer Reliability		81		67	74		72		73	73

of time interval held constant. It may be seen from Table V that the reliability scores in observations using the running record with five and ten categories with time intervals of one minute, thirty seconds, and fifteen seconds were generally higher than the reliability scores using the check sheet record with comparable conditions of observation. However, in observations using the check sheet record and a time interval of five seconds with five and ten categories in the observation, the reliability score was somewhat higher than that for the running record.

With the exception of the one-minute time interval where the reliability score on the check sheet record was slightly higher than that on the running record, the observations using fifteen categories and the running record recording method showed a higher observer reliability score than that of the check sheet record. Reliability was higher with the running record than with the check sheet record when twenty categories were used, with the exception of the five-second time interval where the reliability score on the check sheet record was considerably higher than that on the running record.

Length of Time Interval

Table VI contains the total observer reliability scores for different time intervals with the number of categories and the method of recording held constant. The reliability scores for observations using the running record recording method with five and ten categories decreased consistently with a decrease in the length of the time interval. Observations using the running record and fifteen categories, however, did not show the expected pattern of reliability scores, i.e., there was no pattern of expected consistency as the scores for time intervals of one

TABLE V

TOTAL OBSERVER RELIABILITY SCORES USING THE RUNNING RECORD AND THE CHECK SHEET RECORD WITH NUMBER OF CATEGORIES AND LENGTH OF TIME INTERVAL HELD CONSTANT

Running Record	Obs. 1	Obs. 2	Both Obs.
Group I - 5 Categories			
1 minute	95	100	98
30 seconds	94	95	94
15 seconds	91	89	90
5 seconds	79	91	85
Group II - 10 Categories			
1 minute	91	88	89
30 seconds	85	88	86
15 seconds	87	88	87
5 seconds	72	81	76
Group III - 15 Categories			
1 minute	86	77	83
30 seconds	92	92	92
15 seconds	93	70	83
5 seconds	89	89	89
Group IV - 20 Categories			
1 minute	84	84	84
30 seconds	81	78	80
15 seconds	77	72	74
5 seconds	56	64	60

Check Sheet Record	Obs. 1	Obs. 2	Both Obs.
Group I - 5 Categories			
1 minute	95	83	89
30 seconds	80	91	85
15 seconds	90	89	89
5 seconds	86	94	87
Group II - 10 Categories			
1 minute	88	88	88
30 seconds	81	93	87
15 seconds	71	78	75
5 seconds	81	90	85
Group III - 15 Categories			
1 minute	84	85	84
30 seconds	72	65	67
15 seconds	73	74	74
5 seconds	73	75	74
Group IV - 20 Categories			
1 minute	82	74	79
30 seconds	71	82	77
15 seconds	81	67	74
5 seconds	72	73	73

TABLE VI

TOTAL OBSERVED RELIABILITY SCORES USING DIFFERENT TIME INTERVALS WITH NUMBER OF CATEGORIES AND METHOD OF RECORDING HELD CONSTANT

Running Record

Number of Categories	Length of Time Interval: 1 min.	30 sec.	15 sec.	5 sec.
Group I - 5 Categories				
Obs. 1	95	94	91	79
Obs. 2	100	95	89	91
Both Obs.	98	94	90	85
Group II - 10 Categories				
Obs. 1	91	85	87	72
Obs. 2	88	88	88	81
Both Obs.	89	86	87	76
Group III - 15 Categories				
Obs. 1	86	92	93	89
Obs. 2	77	92	70	89
Both Obs.	83	92	83	89
Group IV - 20 Categories				
Obs. 1	84	81	77	56
Obs. 2	84	78	72	64
Both Obs.	84	80	74	60

Check Sheet Record

Number of Categories	Length of Time Interval: 1 min.	30 sec.	15 sec.	5 sec.
Group I - 5 Categories				
Obs. 1	95	80	90	86
Obs. 2	83	91	89	94
Both Obs.	89	85	89	87
Group II - 10 Categories				
Obs. 1	88	81	71	81
Obs. 2	88	93	78	90
Both Obs.	88	87	75	85
Group III - 15 Categories				
Obs. 1	84	72	73	73
Obs. 2	85	65	74	75
Both Obs.	84	67	74	74
Group IV - 20 Categories				
Obs. 1	82	71	81	72
Obs. 2	74	82	67	73
Both Obs.	79	77	74	73

minute, thirty seconds, fifteen seconds, and five seconds were .83, .92, .83, and .89 respectively. Observations with the running record using twenty categories showed a consistent decrease for each decrease in length of time interval. It should be noted too, that this decrease was much sharper than in the case of fewer categories.

Observations using the check sheet record recording method also appear in Table VI. Using five and ten categories in the observation, there was no consistent pattern of reliability scores, i.e., the reliability scores decreased and then increased with a decrease in the length of the time interval. Observations using fifteen categories show a sharp decrease in reliability from a one-minute to a thirty-second time interval and an increase in the reliability score from a thirty-second to a fifteen-second interval. The reliability score for the fifteen-second and five-second time intervals were identical. Using twenty categories the reliability decreased consistently with a decrease in the length of the time interval.

Number of Categories

Table VII contains the total observer reliability scores for different groups of categories with the length of the time interval and the method of recording held constant. It will be noted that the reliability scores on the running record for a time interval of one minute decreased when the number of categories was increased from five to ten and from ten to fifteen, but that there was a slight increase in reliability when twenty categories were used in the observation.

Observations using the running record and a thirty-second time interval showed a decrease in reliability between five and ten categories and then an increase in reliability between ten and fifteen categories.

TABLE VII

TOTAL OBSERVER RELIABILITY SCORES USING DIFFERENT GROUPS OF CATEGORIES WITH LENGTH OF TIME INTERVAL AND METHOD OF RECORDING HELD CONSTANT

Running Record

Length of Time Interval	Number of Categories 5	10	15	20
1 minute				
Obs. 1	95	91	86	84
Obs. 2	100	88	77	84
Both Obs.	98	89	83	84
30 seconds				
Obs. 1	94	85	92	81
Obs. 2	95	88	92	78
Both Obs.	94	86	92	80
15 seconds				
Obs. 1	91	87	93	77
Obs. 2	89	88	70	72
Both Obs.	90	87	83	74
5 seconds				
Obs. 1	79	72	89	56
Obs. 2	91	81	89	64
Both Obs.	85	76	89	60

Check Sheet Record

Length of Time Interval	Number of Categories 5	10	15	20
1 minute				
Obs. 1	95	88	84	82
Obs. 2	83	88	85	74
Both Obs.	89	88	84	79
30 seconds				
Obs. 1	80	81	72	71
Obs. 2	91	93	65	82
Both Obs.	85	87	67	77
15 seconds				
Obs. 1	90	71	73	81
Obs. 2	89	78	74	67
Both Obs.	89	75	74	74
5 seconds				
Obs. 1	86	81	73	72
Obs. 2	94	90	75	73
Both Obs.	87	85	74	73

A sharp decrease appeared in the reliability of the observations when twenty categories were used. Using a time interval of fifteen seconds, the reliability scores on the running record decreased consistently with a decrease in the length of the time interval.

In the observations using the five-second time interval the reliability of observation scores was lower for ten categories than it was for five. An increase in reliability appeared for fifteen categories, but a sharp decrease in reliability was evidenced in the observations involving twenty categories.

Observations using the check sheet record recording method also appear in Table VII. Using time intervals of one minute, fifteen seconds, and five seconds, the reliability scores on the check sheet record decreased consistently with a decrease in the length of the time interval. Observations using the check sheet record and a time interval of thirty seconds, however, did not show the expected pattern of reliability scores, i.e., there was no pattern of expected consistency as the scores for five, ten, fifteen, and twenty categories were .85, .87, .67, and .77 respectively.

CHAPTER IV

DISCUSSION

The hypothesis guiding this study was that as the number of categories increased and the length of the time interval decreased, reliability of observation would decrease with both a running record and a check sheet record, but would decrease more with a check sheet record.

The results of the study tend to support this hypothesis but not to the degree that was expected when the study was undertaken. The observers were of the opinion that there would be a wide variance in reliability scores, the running record in all cases providing the more reliable method of observation. The results do not bear this out, however, as in some cases the decrease in the reliability scores for the check sheet record was relatively small, and, in others, there was no decrease at all, but rather an increase in the reliability scores.

As predicted, as the number of categories used in the observation was increased and as the length of the time interval used in the observation was decreased the reliability scores for almost all conditions of observation decreased.

Some possible reasons for expected differences in the reliability scores for the various conditions of this experiment were discussed in the introductory chapter. It will be recalled that this discussion pointed to the significance of observer load in reliability of observation. These factors will be discussed here again, in view of the results

of this study, in order to point out with some basis the relative contribution of these different factors to observer load.

From the results of this study it would still appear that observer load is the predominant factor affecting reliability, that is, the more the observer has to do, the lower will be the reliability scores. Evidence supporting this proposition is found in the fact that as additional groups of categories were added to the observer's load, a decrease in reliability generally appeared. The proposition is further supported by the fact that as shorter time intervals were introduced in the observations, the reliability scores also generally decreased.

On the basis of the data in Tables VI and VII it is not possible to determine which of these two factors contributed more significantly to the decrease in reliability of observation, that is, which of the two factors contributed more significantly to the observer's load. It should be noted, however, that the reliability scores computed for the time intervals were subject to two sources of error, that resulting from actual classificatory error due to the increased attention of the observer to the shorter time intervals themselves, and to the greater opportunity for observer error to result from disagreement between observers on the placement of a behavior in its appropriate time interval. The reliability scores computed for the number of categories were subject only to the error resulting from the increase in the observer's load due to the increase in the number of categories used in the observation. It would appear, therefore, that if the reliability scores for the time intervals were corrected for the error attributable to timing errors, they would be higher than the reliability scores obtained for the increase in the number of categories. This implies that perhaps the number of categories that are used in an observation contributes more to observer load than

does the length of the time interval used. This observation is especially interesting in view of the fact that the observers had the impression that the length of the time interval contributed more to their load than did the increase in the number of categories.

A comment would seem to be in order here in regard to the problems encountered in computing the reliability scores for the different time intervals. When observing with the one-minute and thirty-second time intervals, a greater number of interactive behaviors were generally recorded within each interval than when fifteen or five-second time intervals were used. The larger number of categories recorded in a time interval provided a situation that tended to produce errors resulting from lack of agreement between observers in the order of placement of recorded categories within the time interval. For example, the records of two observers observing the same event simultaneously with a running record could take the form of the illustration below.

Figure 1. Illustrating the Conditions that give rise to "Timing Errors" in the Use of the Running Record with a One-Minute Time Interval

In this situation Observer A had recorded an o and an f as his first two entries in the time interval. Observer B had made the same entries in the same order of appearance. Observer A next recorded an o while Observer B recorded an e. This is an error in the records. The point to be emphasized here, however, lies in the fact that all of the

categories appearing in the time interval after this one error were also labeled as errors. This was an arbitrary decision, as a case could be made for considering the records of Observers A and B as in agreement on the order of the last three entries, and therefore, that they probably referred to the same behaviors. Such a consideration would involve surmise, however, and for purposes of this study, it was decided that whenever one or more category entries within one observation record did not coincide identically with the order of appearance within a time interval of the entries within the other observer's record, these category entries would be considered as errors.

A somewhat different problem appears in this connection when dealing with the check sheet record. The difficulty here lies in not being able to identify the sequence in which the categories appeared within the time interval, and therefore in being forced to consider as errors only the checks that appeared in one record that did not have a matching check in the other observer's record. An example of this situation appears below.

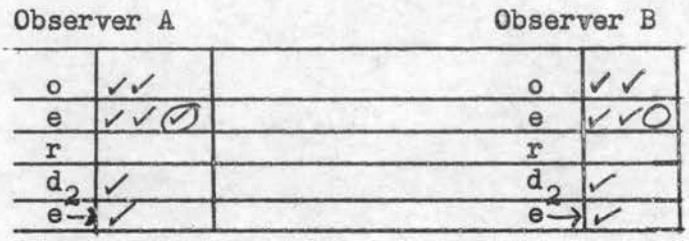

Figure 2. Illustrating the Conditions that give rise to "Timing Errors" in the Use of the Check Sheet Record with a One-Minute Time Interval

In this example Observer A had recorded three checks for category e while Observer B had recorded only two checks for that category. The records in relation to category e would be scored as having two agreements and only one disagreement. One cannot be certain, however, that this third check does not really stand for the same behavior as the second

check that was recorded by Observer B, and therefore, one cannot be certain that the reliability score derived on this basis does not reflect actual error. The effect of this method of determining reliability of observer records may be to spuriously raise the reliability score.

Observations with the fifteen-second and the five-second time intervals usually gave rise to another kind of "timing error" in computing reliability scores. Because of the short intervals, one observer frequently would record a behavior at the end of one interval while the other observer would record the same behavior at the beginning of the following interval. An example of this situation appears below.

Figure 3. Illustrating the Conditions that give rise to "Timing Errors" in the Use of the Running Record with a Fifteen-Second or Five-Second Time Interval

While conditions such as this did not have the possibility of contributing to as many errors in agreement in one interval as did the running record with a one-minute and thirty-second interval, the frequency with which timing errors occurred in the fifteen-second and five-second intervals was greater, and thereby probably contributed a greater number of errors to the total reliability score. Errors of this type that appeared in the fifteen-second and five-second intervals with the check sheet record were handled in a manner similar to that described for the running record.

In connection with the discussion of timing errors it should be pointed out again that any error due to timing was considered as an error only in computing observer agreement, and not in computing the category

reliability scores. This seems a justifiable procedure since in category reliability one is concerned only with the ability of two observers to correctly identify and label interactive behavior, not with the recording of its occurrence in time.

One further fact coming from this study that supports the proposition that observer load is the predominant factor affecting reliability is that, in general, the reliability scores for the observations using the running record were higher than the reliability scores for the observations using the check sheet record. When observing with the running record the categories were committed to memory and the observers simply had to record the appropriate category symbol as the interactive behavior occurred. With the check sheet record, however, it was necessary for the observers to identify the behavior and then record a check on the appropriate blank on the record. This necessitated finding the appropriate category to check and then the appropriate time interval for that category, a recording system considerably more complex than the one involved in the running record. The difference in reliability scores in favor of the running record method of recording evidences this greater complexity. > STOP

There is also good reason to believe that the higher reliability scores for the observations with the running record would have been even greater except for the error introduced into the design of the study through uncontrolled practice effects. It will be recalled that the observers used the categories with the running record first, that is, before the same categories were used with the check sheet record. This provided some two hours of practice with any given set of categories before observations for purposes of data collection with the check sheet record were undertaken. If these practice effects could have been

controlled it is likely that the differences in reliability scores in favor of the running record method of recording would have been greater.

In addition to the contribution of these three factors to observer load there are the factors of observer fatigue and illness that affect the reliability of observation. Observer fatigue may be in part a function of the three factors investigated in this study. However, there are other variables that have their influence, for instance, fatigue from loss of sleep, the length of time that one observes in one day, and the rate of the behavior being observed during that time. Closely related to observer fatigue in terms of effect on reliability scores, is the factor of illness. If an observer is in the early stages of an illness, it is likely that he would not be able to observe as accurately as when he was feeling well.

Besides contributing to the fatigue of the observers, the behavior of the child influences observer reliability in another way, namely, through the problems introduced to accurate measurement by variation in activity level. When the rate or kind of behavior being observed changes sharply, there is good opportunity for observer error, especially if the change is one that involves greater activity or new or unusual behaviors. When the child is displaying behavior that requires many different categories for labeling purposes, or when the child's behavior occurs very rapidly, the observer's load increases proportionately, and observer reliability may decrease proportionately. This problem becomes especially difficult when the observations are being made in terms of short time intervals.

The degree to which the results of this study have been contaminated by the effect of practice on the reliability scores is undetermined. It is evident, however, that the frequently uncontrolled factor of practice

did have an opportunity to affect the reliability scores. The observers had practice in learning the categories initially and again when it was necessary for them to reach an agreement of 80 per cent for each of the categories. Some groups of categories required more practice than others in establishing the 80 per cent agreement, and consequently the observers may have been more familiar with some groups than with others. In addition to this kind of error, the observers were conditioned progressively, so to speak, to the shorter time intervals, i.e., the observers had practice with the fifteen-second time interval prior to the five-second interval and so on, thus making the situation quite unlike the one that would have existed had only the one-minute time intervals been compared with five-second time intervals. It would seem that this factor of practice could have contributed somewhat to the "less than expected" decrease in the reliability scores for decreasing time intervals and for the increased number of categories. Then, too, as indicated earlier, it is believed that the factor of uncontrolled practice contributed to the "less than expected" differences in reliability scores found between the running record and check sheet record methods of recording.

The entire problem of practice effects on observer reliability scores needs further investigation. In addition to some of the problems with respect to practice effects outlined above there is the whole area of question that centers around the maximum number of categories any one person can handle in any one observation. There is some evidence that does indicate that with sufficient practice a large number of categories can be handled in any one observation. For example, Schalock was able to establish reliability for a category system that involved 154 categories, 82 related to mother interactive behavior and 72 to child interactive

behavior (15). It is still not known, however, what the maximum number of categories is that could be reliably handled by any one observer.

A comment might be made about changing subjects during the experiment, and the effect this could have had on the results of the study. It will be recalled that the subject chosen for observation dropped from the nursery school group about half way through the study. The child ranked second (see p. 10) by the nursery school staff was then selected for the remaining observations. The second child was at first very much aware of the presence of the observers, and would often just sit and stare at them. It may be noted that in Table III, p. 24, where observations of the second child began, the results indicate that the child's activity was restricted to a few categories of behavior, and that for the most part the observers were able to demonstrate good reliability on these few categories. Fortunately, this kind of behavior continued for only a few observations. After the initial period of two or three days, the child's behavior closely resembled that of the first child's in both kind and rate. When the behavior of the second child more closely resembled that of the first, the observer reliability scores returned to a pattern that more closely resembled that for the first child. This situation illustrates again the importance of the kind and rate of behavior being observed in reliability measurement.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

The purposes of this study were three:

1. To determine if reliability of observation varied with the method of recording when the number of categories and the length of the time interval used in the recording were held constant.

2. To determine if reliability of observation decreased with a decrease in the length of the time interval when the number of categories and the method of recording were held constant.

3. To determine if reliability of observation decreased with an increase in the number of categories used in the observation when the length of the time interval and method of recording were held constant.

To accomplish these purposes, one three-year-old child enrolled in the Oklahoma Agricultural and Mechanical College Nursery School-Kindergarten Laboratory was selected for observation. Observations for the study were made in terms of a predetermined system of categories. Reliability of observation was demonstrated by two observers who worked together throughout the study. Observer reliability was found by computing the per cent of agreement between the two observers on an item-by-item comparison of their records taken simultaneously but independently. The formula used to calculate the per cent of agreement was:

$$\frac{\text{number of agreements}}{\text{number of agreements plus number of disagreements}}.$$

Observations were made and tested for reliability using a running record and a check sheet record recording method. Practice observations began with the running record, five categories, and a one-minute time interval. After an agreement of .80 for each category had been established, observations for purposes of data collection began. These observations began with the same conditions for which reliability was demonstrated. Both observers observed under these conditions for two fifteen-minute periods of time. When these observations were completed, the time interval was changed to thirty seconds and the observers again observed for two fifteen-minute periods. The same procedure was followed for the fifteen-second and five-second time intervals.

The same five categories were used in observation with the check sheet record. Practice observations using a one-minute time interval were again undertaken. These observations followed the same procedure as that for the running record.

When all of the conditions of observation using five categories were completed, five new categories were added to the original list. The same procedure as that described for five categories was carried out with these ten categories. A similar procedure was followed using fifteen and twenty categories.

Conclusions

From the results of this study, three conclusions seem justified.

1. Reliability scores for all of the conditions of observation were generally higher for the running record than they were for the check sheet record.

2. Reliability of observation generally decreased with a decrease in the length of the time intervals used in the observation.

3. Reliability of observation generally decreased with an increase in the number of categories used in the observation. STOP

These conclusions support the hypothesis that the observer's load is the most important single factor affecting the reliability of observation, and they seem to suggest that the running record method of recording has advantages over the check sheet record method. These results do not indicate, however, the maximum number of categories that can be used in any one observation schedule, the minimum length of time that may be used reliably as a recording interval, or the effects of continued practice on the reliability of observation. Problems such as these await further research.

REFERENCES

1. Arrington, Ruth E. Interrelations in the behavior of young children. Child Develpm. Monogr., 1932, No. 8.

2. Arrington, Ruth E. Time sampling in studies of social behavior: A critical review of techniques and results with research suggestions. Psychol. Bull., 1943, 40, 81-124.

3. Arrington, Ruth E. Time-sampling studies of child behavior. Psychol. Monogr., 1939, 51, 2-193.

4. Barker, M. A technique for studying the social-material activities of young children. Child Develpm. Monogr., 1932, No. 7.

5. Beaver, A. P. The initiation of social contacts by preschool children. Child Develpm. Monogr., 1932, No. 7.

6. Goodenough, F. L. Inter-relationships in the behavior of young children. Child Develpm., 1930, 1, 29-47.

7. Goodenough, F. L. Measuring behavior traits by means of repeated short samples. J. juv. Res., 1928, 12, 230-35.

8. Heyns, Roger W., and Lippitt, R. Systematic observational techniques. In Gardner Lindzey (Ed.), Handbook of social psychology. Combridge, Mass: Addison-Wesley, 1954, Vol. 1, Pp. 370-404.

9. Loomis, A. M. A technique for observing the social behavior of nursery school children. Child Develpm. Monogr., 1931, No. 5.

10. Murray, E. J. A case study in a behavioral analysis of psychotherapy. J. abnorm. soc. Psychol., 1954, 49, 305-310.

11. Olson, Willard C. The measurement of nervous habits. In Wayne Davis (Ed.), Readings in child psychol. 1951, Pp. 439-45.

12. Parten, M. B. Leadership among pre-school children. J. abnorm. soc. Psychol., 1933, 27, 430-40.

13. Parten, M. B. Social participation among pre-school children. J. abnorm. soc. Psychol., 1932, 27, 243-69.

14. Parten, M. B. Social play among pre-school children. J. abnorm. soc. Psychol., 1933, 28, 136-47.

15. Schalock, Henry D. Observation of mother-child interaction in the laboratory and in the home. Unpublished PhD. Dissertation, Univ. of Nebraska, 1956.

16. Thomas, D. S., Loomis, A. M., and Arrington, R. E. Observational studies of social behavior. Vol. 1. Social behavior patterns. Institute of Human Relations, Yale University, 1933.

17. Thomas, D. S., et al. Some new techniques for studying social behavior. Child Develpm. Monogr., 1929, No. 1.

VITA

Jo Ann Budzik

candidate for the degree of

Master of Science

Thesis: RELIABILITY OF OBSERVATION AS A FUNCTION OF RECORDING PROCEDURE, NUMBER OF CATEGORIES, AND LENGTH OF TIME INTERVAL: A METHODOLOGICAL STUDY

Major: Family Relations and Child Development

Biographical:

Personal data: Born in Columbus, Ohio, May 14, 1934.

Education: Attended grade school at Holy Family in Tulsa, Oklahoma; graduated from Holy Family High School in May, 1951; received the Bachelor of Science degree from the Oklahoma Agricultural and Mechanical College, with a major in Home Life, in May, 1955; completed the requirements for the Master of Science degree in May, 1956.

Professional experience: Graduate Assistant for Department of Family Relations and Child Development, 1955-1956.