8-1971

Observer Validity

Riley C. Worthy
Kansas State College

OBSERVER VALIDITY

A Thesis Submitted to the Graduate
Department in Partial Fulfillment of
the Requirements for the Degree
of Master of Science

by

Riley C. Worthy

Kansas State College of Pittsburg
Pittsburg, Kansas
August 1971

ACKNOWLEDGEMENTS

The author wishes to thank the following persons whose support and encouragement made this thesis possible: Dr. Sebastian Striefel and Dr. Robert Fulton who made this thesis possible; Drs. Vance Cotter and Joseph Spradlin whose sound advise and criticism were of inestimable value and to Dr. Herbert Rumford whose tolerance was deeply appreciated; and finally to Mrs. Ruth Staten and Mrs. Candy Dear whose typing skills are exceeded only by their gracious cooperation.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

The accuracy of data recorded by eight different observers over five observational tasks was investigated. The measures collected by each observer were compared with criterion values for each task. In addition, a coefficient of reliability was found for each observer by comparing his performance with that of the group. The results showed that the accuracy of observer recorded data was within the limits of acceptability and that the correlation between reliability and accuracy was moderately high.

CHAPTER I

INTRODUCTION

The research methods of the experimental analysis of behavior stress the precise definition and measurement of observable behavior. The process of determining change becomes the process of counting behaviors automatically or with human observers. In the case of human observation, accuracy may vary over time depending on motivation, previous training, session duration and target frequency, intensity, and duration.

Although the history of psychology places observer error in a class by itself, published reports of research involving observer recorded data rarely evaluate the accuracy of observation other than in terms of inter-observer reliability. Whether the observer's records are valid measures of the specified behavior is a question which remains essentially unanswered.

During an experimental investigation of the educational uses of television in progress at Parsons State Hospital and Training Center, an effort was made to assess the visual attention holding properties of various portions of the "Sesame Street" educational telecast upon retarded children. Observers, viewing children through one-way mirrors, were required to press a button when a child was not watching the program. The observers responses were then recorded in terms of the number of responses made and the total response duration for each observational

period. Observer accuracy was evaluated after each session by measuring the extent to which observers agreed with each other. The measures of inter-observer reliability were acceptably high, when averaged over a period of several months, but daily variations suggested that the observers data might be inaccurate and that the observer mode of data collection be inappropriate for this class of behavior.

I. THE PROBLEM

Statement of the Problem

This thesis has two objectives. The first is to assess the accuracy of the observer mode of recording absolute values of frequency and duration of television watching.

The second objective is to demonstrate the relationship between measures of observer accuracy and measures of inter-observer reliability.

Need for the Study

The value of this study is in its ability to indicate the level of confidence which may be attributed to observer collected data under conditions similar to those specified in this study.

Limitations

(1) The participants in this study were required to monitor a videotaped presentation of one person watching television as opposed to a "live" presentation of a group of children doing the same thing.

(2) Simultaneous recording of the experimental stimulus and the criterion on videotape is a critical variable. Although carefully recorded and edited by both the experimenter and the subject of the tape it is estimated that a duration error of about one percent exists.

(3) The television watcher in this experiment was a normal adult rather than a retarded child.

Delimitations

(1) Any assessment of observer accuracy is limited to the nature of the observational task. The level of observer accuracy reported in this study is specific to the television watching behavior

of retarded children.

(2) The relationship between observer reliability and observer accuracy as reported in this study is delimited to the size of the observer group and to the method of computing the reliability coefficient.

Definition of Terms

(1) Validity. Validity is defined as the accuracy with which observers are able to record the criterion behavior. Within the context of this thesis the terms validity and accuracy are synonomous.

(2) Reliability. Reliability refers to inter-observer reliability and is defined as per cent of agreement between independent observers.

(3) Target Behavior. The behavior which the observers are expected to record. It is the frequency and duration that a television watcher looks away from or breaks eye contact with the television screen.

(4) Performance Scores. Derived scores which express the observers agreement with the criterion. The scores expressing agreement between the observer's measure and the actual duration (the criterion value) is duration accuracy (DUR. ACC.). The scores expressing agreement between the observers measure and the actual frequency of the criteria is frequency accuracy (F. ACC.).

Hypotheses

(1) That there will be no significant difference between performance scores for each observation session.

(2) That there will be no significant difference between observers.

(3) That there will be no significant correlation between performance scores and measures of inter-observers reliability.

Organization of Remainder of Thesis

Chapter II will present a review of the literature central to the problem. Chapter III will present the method, and procedures used in the problem. Chapter IV will present the results and Chapter V will present the discussion and recommendations.

CHAPTER II

REVIEW OF THE LITERATURE

The task of the observer, in addition to identifying the required behavioral event, is to provide a permanent record of it with pencil and paper or with the electromechanical assistance of automatic recording apparatus. Electromechanical recording has certain advantages over the paper and pencil methods. The former requires less attention than the latter, thus allowing the observer to devote more of his efforts to observing.

The electromechanical methods are described elsewhere,[1,2,3] but are all essentially the same in that the observers task is to press a key when an event occurs. The keys control corresponding counters or other classes of read-out facilities at a remote location.

---

[1] O. I. Lovaas, et al., "Experimental Studies in Childhood Schizophrenia: Analysis of Self Destructive Behavior," Journal of Experimental Child Psychology, Vol. 2, (1965), p. 70 (a).

[2] O. I. Lovaas, et al., "Recording Apparatus and Procedure for Observation of Behaviors of Children in Free Play Settings," Journal of Experimental Child Psychology, Vol. 2, (1965), pp. 108-120, (b).

[3] R. C. Wahler, et al., "Mothers as Behavior Therapists for their own Children," Behavior Research and Therapy, Vol. 3, (1965), pp. 113-124.

There are three styles of recording events in field situations: one consists of logging the incidences of responses (and, in many cases, their durations);[4,5] the second consists of registering the frequencies of occurrence and non-occurrence within a time interval;[6] the third and most cumbersome style is to record frequency and duration at the same time.[7,8] The last two styles are often employed in the field of behavior modificiation where rate of response and total duration are of primary interest.

Since observational methods are usally employed where more objective instrumentation is inappropriate or impossible the adequacy of the data is hard to evaluate. One could hardly expect that accurate reporting or recording would follow as an automatic consequence of observing. That it does not is clearly indicated by the amount of disagreement which often appears in reports of the same event by different observers.

---

[4] B. M. Hartz, et al., "Effects of Social Reinforcement on Operant Crying," Journal of Experimental Child Psychology, Vol. 1, (1964), pp. 145-153.

[5] R. P. Hawkins, et al., "Behavior Therapy in the Home: Amelioration of Problem Parent-Child Relations with Parent in a Therapeutic Role," Journal of Experimental Child Psychology, Vol. 4, (1966), pp. 99-107.

[6] F. R. Harris, et al., "Effects of Positive Social Reinforcement on Regressed Crawling of a Nursery School Child," Journal of Educational Psychology, Vol. 55, (1964), pp. 35-41.

[7] Lovaas, (1965a), Op. cit., pp. 67-84.

[8] K. E. Allen, et al., "Effects of Social Reinforcement on Isolate Behavior of a Nursery School Child," Child Development, Vol. 35, (1964), pp. 511-518.

Reliability of observation has generally been defined in observational studies as the percentage of agreement between two observers who have recorded the same behavior at the same time.[9] In the experimental laboratory this is accomplished by occasionally bringing in a second observer. Both observers then record the target behavior simultaneously and the observers are prevented from interacting with each other either by placing a barrier or an appropriate amount of space between them. This way the investigator has two independent records of the same behavior which he may compare in order to get a measure of reliability.

The reliability index is calculated from the records produced by each observer. The frequency or duration of the observed phenomena, which ever the case may be, is totaled for each of the observers and the smaller total is divided by the larger. If both totals are identical the reliability index would be 1.0.[10] For example, if observer A recorded 10 behavioral events and observer B recorded 12 behavioral events, the reliability index would be equal to 10 divided by 12, or .83. This method is often used when the investigator is interested in frequencies per se, since the measure

---

[9]Herbert F. Wright, "Observational Child Study," Handbook of Research Methods in Child Development, Paul H. Mussen (Ed.), (New York,

[10]Bijou, et al., "A Method to Integrate Descriptive and Experimental Field Studies at the Level of Data and Empirical Concepts." Journal of Applied Behavior Analysis, Vol. 1, (1968), pp. 183-184.

obtained gives only the amount of agreement over the total number of events observed. It does not indicate whether the two observers were recording the same event at the same time. Thus, it might be possible that one observer was recording few behaviors during the first half of the session and many during the second, while the second observer was doing just the opposite. To ascertain whether this is the case, one could divide the period of observation into small segments and calculate the reliability of each. Agreement over progressively smaller segments, such as 5 or 10 seconds, gives confidence that the observers are scoring the same event at the same time. Reliability is calculated by scoring each interval as agree or disagree (match or mismatch) and dividing the total number of agreements by the number of agreements plus the number of disagreements.[11]

Using the procedures just described a study was undertaken at the University of Illinois to obtain a descriptive account of a boy in a laboratory nursey school.[12] The behaviors recorded were social contacts and sustained activities. Observation began four weeks after the start of the school year and covered a three hour period in the morning. Observations were taken on 28 school days. Every 10 seconds the teacher recorded, with pencil and paper, occurrence or non-occurrence of the target behavior. Reliability of observation was evaluated several times throughout the study by having a second observer record

---

[11] Ibid

[12] Bijou, Op. cit., pp. 185-190.

the behavior. Reliability was calculated by scoring each interval as a match or mismatch and dividing the total number of agreements by the number of agreements plus disagreements. Four checks on social contacts yielded agreements of 75, 82, 85 and 87%. Three checks on sustained activity showed agreements of 94, 95, and 97%. Thus, average agreement on social contacts exceeded 82% while average agreement of sustained activity exceeded 95%.

An automated observational system using a stenograph was reported by Boer.[13] The study was an attempt to provide a descriptive account of the free play activities of four disturbed children between the ages of four and five. Eight behavior categories were defined and included such gross behavior as playing with another child, playing with adults, playing with toys, locomotion and resting. Each child was observed for five minutes each morning over a period of 5 weeks. The observer, a secretary without previous training in behavioral research or scientific methodology, was trained by the author to identify the eight categories of behavior and then to practice the recording technique under close supervision for three 15 minute sessions on three consecutive days. Each key of the stenographic machine represented a given category of behavior and the observers task was to press the key at one second intervals, as marked by the beat of an electrical metronome, whenever the child engaged in the behavior. Recordings were made on paper tape that advanced each

[13] Arand P. Boer, " Application of a Simple Recording System to the Analysis of Free-Play Behavior in Autistic Children," Journal of Applied Behavior Analysis, Vol. 1, (1968), pp. 335-340.

time a key was pressed. On the fourth training day the observer recorded the behavior of the four children independently. At the same time, the investigator recorded the behavior of each child with another stenograph and found 90% agreement between himself and the observer. Subsequent checks on reliability were determined for days three, six and nine by recording the childrens play on closed circuit television and, six weeks later, having the observer re-record the behavior from the video tape. Agreement between the original and play-back observations by the same observer was of 95%. Two other observers, similarly trained and evaluated showed agreements of over 93%.[14]

Comparing the performance of observers does not insure accuracy of recording. Bijou[15] points out that both observers might record some events which should be noted and ignore others which should. One such instance is reported by Bernhardt[16] in a study of social contacts by children with their peers in a nursey school. A motion picture camera recorded each subject while paired observers recorded the target behavior. The same observers then recorded the same behavior from the motion picture film. The film was projected repeatedly until differences between the observers were resolved. Although the per cent of agreement between the two observers was 82, the number of observed

---

[14] Boer, Op. cit., p. 337.

[15] Bijou, Op. cit., p. 185.

[16] K. S. Bernhardt, et al., "An Analysis of the Social Contacts of Preschool Children with the Aid of Motion Pictures," Toronto University Studies; Child Development Series, No. 10, (1937), Toronto: University of Toronto Press.

social contacts recorded was increased approximately 70%[17] by addition of what was seen on the screen to what was seen in the field.

Reliability of observation has been examined extensively and reported measures of observer agreement in different studies do not clearly favor any principle method over any other. Notably absent in the literature of behavior modification and child developement are systematic efforts to assess the accuracy of observer collected data for the samller, more literally objective, operant classes of behavior. Heyns and Lippitt[18] sum it up quite eloquently: "It has been pointed out by Cronbach (1946) and Guttman(1950), and others that the term *validity* has a variety of meanings. In one sense, the validity question means: Does the observer score measure what it purports to measure? In another sense, the question of validity asks: Does the observer score predict anything? If we accept the first meaning of the term, the question of validity of observer scores has been relatively ignored by those people most active in using observers."

---

[17]Bernhardt, Op. cit., p. 11.

[18]Roger W. Heyns, Ronald Lippitt, "Systematic Observational Techniques," Handbook of Social Psychology, Gardner Lindzey (Ed.), (Cambridge, Addison-Wesley, 1954), pp. 397-398.

CHAPTER III

METHOD AND PROCEDURES

This study was carried out by presenting a videotaped recording of a person watching television, to eight observers who were told to record the frequency and duration of non-television watching.

I. SUBJECTS

Six female and two male research assistants between the ages of 20 and 40 were recruited from the staff of the Bureau of Child Research at Parsons State Hospital and Training Center, to serve as observers. Five of the assistants had previous experience either as observers or monitors in psychological or psychoacoustic research. The three remaining were primarily clerical personnel with limited experience either as observers or monitors.

II. APPARATUS

Experiment Room

The experimental room was a partially sound attenuated room located adjacent to a control room containing programing equipment. The ambient noise level of the room was 60 decibels (plus or minus five decibels) as measured by a sound level meter. The room was darkened for comfortable video monitoring and was adequately ventilated and air conditioned. A

television monitor was placed at eye level at one end of the room. Observers were located eight feet from the monitor and isolated from each other by a fiber panel. Externally generated noises were masked by presenting 65 DB of white noise into the experimental room via a speaker adjacent to the television monitor.

The Videotape Presentation

The videotape presentation consisted of two channels of information; the visual presentation of the experimental stimulus on one channel and criterion data recorded on an adjacent channel.

Experimental Stimulus

The experimental stimulus was a visual presentation of a person watching television. The view presented provided essentially the same view offered observers participating in a study of educational uses of television; a frontal view of the watcher and the back of the television set being watched, all from a distance of about 8 to 12 feet. The television watcher, i. e., the stimulus-subject, was a normal adult rather than a retarded child, who displayed episodes of non-watching behavior on a pre-arranged schedule. The display was arranged during the videotaping process. The television watcher was told to look away from the television when cued by an audible tone and to look back when the cue ended. The cue frequencies and durations were produced from previous records of patients viewing "Sesame Street". The schedule of these cues was reproduced on a tape recorder and played back to the television watcher during the videotaping process, thus replicating the target behavior of patients. For each of the five

presentations the records of a different patient were used to cue the stimulus-subject.

Crtierion

A 1000 cycle tone was recorded on the second channel of the videotape for the duration of each episode of non-watching behavior. The tones were recorded by the television watcher who pushed a switch when he looked away from the television set and released it when he looked back. These series of tones were the criterion measures of the target behavior, and were used in the data recording process.

Data Recording

The programing apparatus and the videotape recorder were located in a room adjacent to the experimental room. Observers' responses, and criterion information from the second channel of the videotape, were fed into programing apparatus which processed the data and recorded it on a four channel audio-tape recorder. Six kinds of information were recorded:

(1) The total duration, in seconds, that the observer held the response key closed. This was the observed duration measure (OBS. DUR.).

(2) Duration, in seconds, that the observer held the response key closed in the presence of the criteria. This is the amount of time that the observers correctly identified the criterion duration (COR. DUR.).

(3) The observed frequency (OBS. F.); the number of times that the observer pressed the response key.

(4) The number of times that the observer pressed the response

key, in the presence of the criterion. This is the number of times that the observer correctly identified the occurence of the target behavior (COR. FREQ.).

(5) The number of times the criterion occurred (CRIT. F.).

(6) The duration in seconds that the criterion was on (CRIT. DUR.).

Performance Scores

Each observer was given a score for the accuracy with which he recorded the duration and frequency of the criterion behavior. Scoring was considered essential because the raw scores alone did not reflect all of the errors that the observer made in recording the behavior. For example, from the duration data in Appendix A, for the first session, subject A reported the target duration to be 214.9 seconds, but, of those 214.9 seconds, Appendix B shows that only 194.5 seconds were recorded when the criterion was present. The observer recorded 20.4 seconds of target which did not exist. The observer made an error of commission, he pushed the button when he shouldn't have.

Another class of error common to all observers is the error of omission. Considering observer A again, Appendix B and E show that of the 202.2 seconds of target behavior which were actuallty presented during the first session, the observer recorded only 194.5. The observer committed an error of omission in failing to record the remaining 7.7 seconds of target behavior. The sum of observer A's error of omission and omission total 28.1 seconds.

Performance scores, based on criterion values, and reflecting observer errors, were computed for each observers accuracy in recording both frequency and duration:

$$\text{Accuracy for Duration (DUR. ACC.)} = \frac{\text{Criterion Duration}}{\text{Criterion Duration + Sum of Errors}} \times 100$$

where,

Errors of Commission = Observed Duration - Correct Duration

and,

Errors of Ommission = Criterion Duration - Correct Duration

The accuracy score for frequency was similarly computed:

$$\text{Accuracy for Frequency (FREQ. ACC.)} = \frac{\text{Criterion Frequency}}{\text{Criterion Frequency + Sum of Errors}} \times 100$$

where,

Errors of Commission = Observed Duration - Correct Frequency

and,

Errors of Ommission = Criterion Frequency - Correct Frequency.

The performance scores thus selected are independent of observed duration, have a value of 1.0 for perfect accuracy, and a value of 0.5 when criterion measure and error measure are equal.

III. PROCEDURE

Each observer was scheduled to participate as his work load permitted, with a minimum interval of one hour and a maximum of one day between 10 minute trials. All subjects completed the project within

four days with each session accommodating two subjects.

Instructions were given at the beginning of every session. Each subject was told that he would be observing a person watching a television program, and that he was to record when the person was not watching. The observer was given a hand switch and told that he was to record all "non-watching" behavior by depressing the button at the onset of the behavior and by releasing at its cessation. Questions of clarifications were answered by reiterating the instructions in essentially the same form or by stating that further explanation would be made available at the end of the study.

CHAPTER IV

RESULTS

In an examination of the accuracy of human observation, eight subjects, acting as observers, monitored five videotaped presentations of a person watching a television program. For each subject, both the number and duration of responses were recorded. From this data each subject was scored for his accuracy in recording the target behavior. This chapter presents the statistical analyses of these scores. A level of confidence of .05 was set as a criterion for evaluating the results of these analyses.

I. Effects Between Trials

Before an analysis of observer accuracy could be carried out it was necessary to show that there were no systematic effects between the five treatments, i.e., the five videotaped presentations. For an analysis of repeated measures on the same individuals, a trend analysis[1,2] for stability of measures was computed for the accuracy scores in Tables I and II. The results of the trend analysis are shown in Table III. Inspection of the Table reveals an 'F' ratio of 2.54 for the duration scores and

---

[1] Allen L. Edwards, Experimental Design in Psychological Research, Holt, Rhinehart-Winston, New York, (1960), pp. 224-227.

[2] James L. Brunning and B. L. Kintz, Computational Handbook of Statistics, Scott, Foresman & Co., Glenview, Ill., (1968), pp. 42-47.

TABLE I

DURATION ACCURACY SCORES, MEANS, AND STANDARD DEVIATIONS

	Session						
Subject	I	II	III	IV	V	Mean	SD
A	87	90	92	80	88	87	4.6
B	89	89	90	84	76	86	5.9
C	90	92	93	88	85	90	3.2
D	91	93	93	89	90	91	1.8
E	93	92	94	89	90	92	2.1
F	89	89	81	85	84	86	3.5
G	87	88	89	87	87	88	1.0
H	72	88	89	86	87	84	7.0
Mean	87	90	90	86	85		
SD	6.5	2.0	4.15	3.0	4.5		

Total Mean = 87.9, SD = 4.50, Range = 72 to 94

TABLE II

FREQUENCY ACCURACY SCORES, MEANS

AND STANDARD DEVIATIONS

	Session						
Subject	I	II	III	IV	V	Mean	SD
A	84	91	82	73	84	83	6.5
B	84	96	82	73	76	82	8.9
C	87	88	82	82	91	86	3.9
D	95	91	90	89	84	90	4.0
E	87	88	86	86	84	86	1.5
F	80	91	90	86	84	86	4.5
G	75	81	82	80	86	81	4.0
H	37	55	63	61	64	56	11.2
Mean	79	85	82	79	82		
SD	17.8	12.9	8.5	9.3	8.2		

Total Mean = 81.25, SD = 12.59. Range = 37 to 96

TABLE III

SUMMARY OF TREND ANALYSIS FOR FREQUENCY

AND DURATION ACCURACY SCORES

FREQUENCY

Source	SS	df	Ms	F	P
Total	5190	39			
Subject	3929	7	561.28	15.28 **	$<.001$
Sessions	233	4	58.25	1.58 *	$>.05$
Error	1028	28	36.71		

DURATION

Source	SS	df	Ms	F	P
Total	793	39			
Subject	253	7	36.14	2.55**	$<.05$
Sessions	144	4	36.0	2.54*	$>.05$
Error	396	28	14.14		

* The value of 'F' required for 4 and 28 degrees of freedom at the .05 level of significance is 2.71.

** The value of 'F' required for 7 and 28 degrees of freedom at the .05 level of significance is 2.36.

1.58 for frequency. Both figures fall below the 2.71 value required for significance at the .05 level for 4 and 8 degrees of freedom. Since the 'F' ratios were insignificant, the first hypothesis presented in chapte one, that there would be no significant difference between performance scores in each session, was accepted as true. It was concluded that there were no systematic performance effects from one videotaped presentation to the next.

II. Differences Between Observers

Accurate observation precludes individual differences. If observers recording the same event produce different measures of that event then someone is measuring inaccurately. The eight subjects in this study observed the same videotaped samples of target behaviors. This section presents the analysis of the difference between subjects. The trend analysis employed in the previous section showed that there were no differences in performance from session to session. The same analysis shows that the between subjects difference is significant for both the duration and frequency accuracy scores. The results of the trend analysis between subjects for the performance scores in Tables I and II are shown in Table III. Inspection of Table III shows a between subjects 'F' ratio of 2.55 for duration and 15.28 for frequency. The value of 'F' required for significance at the .05 level for 7 and 28 degrees of freedom in 2.36. The duration accuracy scores are significant at the .05 level while the frequency accuracy scores are significant beyond the .001

level. Since both 'F' values are significant the second hypothesis presented in Chapter one, that there would be no significant difference between observers, is rejected and it is concluded that at least one subject is an inaccurate observer.

III. Correlation Between Reliability and Accuracy

When observers are used to record data in the experimental laboratory or in free field studies it is rarely possible to obtain a direct measure of the observers ability to record data accurately. Observer reliability, on the other hand, is easily measured and to do so is standard operating procedure. If measures of observer reliability and measures of observer accuracy are correlated with each other, it follows that one can be predicted from the other to the extent that they are correlated. This section presents the correlation data for reliability and accuracy, and tests hypothesis number three which was presented in Chapter I.

To test hypothesis number three, reliability coefficients were computed for each observer. The coefficient expresses percent of agreement between the observers raw data and the group mean. The reliability coefficients were computed from the following formula and are presented in Tables IV and V.

$$\text{Reliability for OBS. DUR.} = \frac{\text{OBS. DUR.}}{\text{Session Mean OBS. DUR.}}$$

$$\text{Reliability for Frequency} = \frac{\text{OBS. F.}}{\text{Session Mean OBS. F.}}$$

TABLE IV

OBSERVER RELIABILITY SCORES FOR FREQUENCY

	Session					
Subject	I	II	III	IV	V	x
A	87	91	98	93	87	91
B	87	82	79	93	83	85
C	83	93	98	94	93	92
D	73	91	98	87	90	88
E	83	93	88	89	87	88
F	90	91	88	89	87	89
G	97	98	92	97	95	96
H	51	66	71	78	76	68
x	81	88	89	90	87	

Total Mean = 87.15, SD = 8.54, Range = 51 to 98

TABLE V

OBSERVER RELIABILITY SCORES FOR DURATION

	Session					
Subject	I	II	III	IV	V	Mean
A	92	99	98	90	95	95
B	96	98	99	96	91	96
C	96	99	98	96	96	97
D	96	98	99	98	96	97
E	97	98	99	97	99	98
F	98	98	99	99	98	98
G	97	96	97	95	96	96
H	74	98	97	98	97	93
Mean	93	98	98	96	96	

Total Mean = 96.3, Total SD = 4.17, Range = 74 to 99

A product-moment correlation for the duration accuracy scores in Table I and the reliability coefficients of Table V produced a r of .73. For the frequency accuracy scores in Table II and the reliability coefficients of Table IV, a correlation of .59 was found. Both coefficients are well beyond the value of .312 required for the .05 level of confidence for 38 degrees of freedom. Both correlations are, in fact, beyond the level required for the .01 level of confidence. Since both figures are significant, hypothesis number three presented in Chapter one, that there would be no significant correlation between measures of reliability and measures of accuracy, is rejected.

CHAPTER V

DISCUSSION, RECOMMENDATIONS AND SUMMARY

I. DISCUSSION

Previous research using human observers to measure experimental effects, base the validity of treatment effects on percentage of agreement between two or more observers. The present study developed a technique to compare the accuracy of the human observer with a validity criterion. The results of this study pointed out that an observer was relatively accurate in recording absolute values of duration of non-television watching. Also, seven observers were relatively accurate for frequency of non-television watching. Only one of the eight observers (H) was relatively lower in frequency accuracy. This indicates that the majority of observers' performance was accurate--well within the range of acceptability. Furthermore, the reliability scores indicate that the percentage of agreement between the eight observers was high for duration and frequency. All scores were within the range of acceptability.

Consider the stability of observer performance over sessions. The analyses of variance yielded no significant difference in frequency or duration scores. Such a finding is important in the analysis of behavior because it leads to confidence in validity of behavioral measurement.

II. RECOMMENDATIONS

The validation procedure used in the study appears to be unique in the field of experimental research using human observers. Such a procedure can be used to evaluate or screen observers. For example, inspection of any two of the eight observers' scores yields high accuracy and reliability indices (with the exception of the frequency accuracy score for Subject H). Nevertheless, knowledge of this information permits the researcher to either train the unreliable observer or replace him with a more reliable one.

The results also suggest that one method of screening observers on basis of accurate performance is to compare frequency and duration performance. Inspection of standard deviations and ranges for frequency and duration accuracy scores (Tables I and II) indicate greater dispersion for frequency scores than for duration scores. Furthermore, the reliability scores (Tables III and IV) are lower for frequency and the range is greater. This suggests that analysis of frequency data may be a more sensitive index by which to discriminate differences in observer performance.

III. SUMMARY

Eight subjects, recruited from the observer pool of the Bureau of Child Research at Parsons State Hospital and Training Center, participated as observers in a test of their ability to record behavior accurately. Each observer monitored five, ten minute long videotaped

presentations of a person engaged in watching a televised broadcast of "Sesame Street". Observer accuracy was quantified by a unique method of dividing a videotape recording into channels which simultaneously presented the independent variable to the observer and the criterion variable to programing apparatus so that observer's response and the criteria measure could be compared and recorded at the same time. The results show that observers recorded duration more accurately than frequency, yet both measures were judged valid and reliable.

APPENDIX

APPENDIX A

OBSERVED DURATION (OBS. DUR.)

Total Time in Seconds that Each Observer Held Response Key Closed. This is the observers record of the duration of the target duration.

	Session					
Observer	I	II	III	IV	V	Total
A	214.9	301.1	216.4	296.7	199.0	1228.1
B	206.4	306.7	211.2	257.7	173.1	1155.1
C	302.4	208.1	208.1	258.3	182.2	1158.1
D	207.1	306.1	210.3	273.5	196.0	1193.0
E	204.0	298.4	211.4	261.6	189.7	1165.1
F	202.6	308.6	211.0	269.0	186.9	1178.1
G	203.7	291.9	218.2	257.3	196.6	1167.7
H	148.8	307.2	210.3	273.2	194.1	1133.6
Mean	199.3	302.8	212.1	268.4	189.7	

APPENDIX B

CORRECT DURATION (COR. DUR.)

Total Time in Seconds that Observers Held

Response Key Closed When Criterion was Present

	Session					
Observer	I	II	III	IV	V	Total
A	194.5	284.4	203.3	247.3	181.5	1111.0
B	192.9	285.4	199.0	236.3	152.5	1066.1
C	193.9	289.7	202.3	244.4	170.1	1100.4
D	195.6	292.3	202.6	253.0	182.4	1125.9
E	195.5	286.9	205.0	247.5	179.2	1114.1
F	191.0	286.3	201.1	243.8	171.2	1093.4
G	189.0	275.6	201.8	242.0	179.7	1088.1
H	138.1	284.7	198.1	248.2	177.9	1047.0
Mean	186.3	285.7	201.65	245.3	173.3	

APPENDIX C

OBSERVED FREQUENCY (OBS. F.)

Number of Times the Observers Closed the Response Key

	Session					
Observer	I	II	III	IV	V	Total
A	25	34	21	45	30	155
B	25	32	17	45	41	160
C	24	35	21	40	32	152
D	21	34	21	37	31	144
E	24	35	19	38	30	146
F	26	34	19	38	30	147
G	28	38	23	41	36	166
H	56	56	30	54	45	241
Mean	28.6	37.2	21.4	40.2	34.4	

APPENDIX D

CORRECT FREQUENCY (COR. FREQ.)

Number of Times the Observer Closed

the Response Key While the Criterion was Present

	Session					
Observer	I	II	III	IV	V	Total
A	23	33	17	33	26	132
B	24	31	15	17	33	140
C	23	34	17	37	29	140
D	20	33	19	34	26	132
E	23	33	16	34	26	132
F	23	32	17	35	26	133
G	24	37	20	37	31	149
H	42	40	20	39	27	168
Mean	25.2	34.1	17.6	35.8	28.0	

APPENDIX E

CRITERION DURATION (CRIT. DUR.) IN SECONDS

This Value is the Total Duration of the 1000 Cycle Tone

Session					Total
I	II	III	IV	V	
202.2	299.7	210.0	263.5	189.3	1164.7

CRITERION FREQUENCY (CRIT. F.)

These Values Show how Often the 1000 Cycle Tone was Delivered

Session					Total
I	II	III	IV	V	
21	31	19	33	32	136

BIBLIOGRAPHY

BIBLIOGRAPHY

Allen, K. E., Hart, B. M., Buell, J. S., Harris, F. R., and Wolf, M. M. "Effects of Social Reinforcement on Isolate Behavior of a Nursery School Child." Child Development, Vol. 35, (1964), pp. 511-518.

Arrington, Ruth E. Interrelations in the Behavior of Young Children. Columbia University Press, New York (1932), (Child Development Monograph, No. 8).

Becker, W. C., Madsen, C. H., Jr., Arnold, C. R., and Thomas, D. R. "The Contingent Use of Teacher Attention and Praise in Reducing Classroom Behavior Problems." Journal of Special Education, Vol. 1, (1967), pp. 287-307.

Bijou, S. W., Peterson, R. F., and Ault, M. "A Method to Integrate Descriptive and Experimental Field Studies at the Level of Data and Empirical Concepts." Journal of Applied Behavior Analysis, Vol. 1, (1968), pp. 175-191.

Bijou, S. W., Peterson, R. F., Harris, F. R., Allen, E. K., and Johnston, M. S. "Methodology for Experimental Studies of Young Children in Natural Settings." Psychological Record, Vol. 19, (1969), pp. 177-210.

Bishop, Barbara, M. "Mother and Child Interaction and the Social Behavior of Children." Psychological Monograph, Vol. 65, No. 11 (1951).

Bruning, James L., and Kintz, B. L. Computational Handbook of Statistics. Scott, Foresman & Company, Glenview, Illinois (1968), pp. 43-47.

Edwards, Allen, L. Experimental Design in Psychological Research. Holt, Rinehart & Winston, New York (1960), pp. 224-227.

Green, Elsie H. "Friendships and Quarrels Among Preschool Children." Child Development, Vol. 4, (1933), pp. 237-252.

Harris, F. R., Johnston, M. K., Kelley, C. S., and Wolf, M. M. "Effects of Positive Social Reinforcement on Regressed Crawling of a Nusery School Child." Journal of Educational Psychology, Vol. 55, (1964), pp. 35-41.

Hart, B. M., Allen, K. E., Buell, J. S., Harris, F. R., and Wolf, M. M. "Effects of Social Reinforcement on Operant Crying." Journal of Experimental Child Psychology, Vol. 1, (1964), pp. 145-153.

Hawkins, R. P., Peterson, R. F., Schweid, E., and Bijou, S. W. "Behavior Therapy in the Home: Amelioration of Problem Parent-Child Relations with Parent in a Therapeutic Role." Journal of Experimental Child Psychology, Vol. 4, (1966), pp. 99-107.

Jensen, G. D., and Bobbitt, Ruth A. "Implications of Primate Research for Understanding Infant Development." The Exceptional Child, J. Hellmouth, editor. Special Child Publications, Seattle, Washington, Vol. 1, (1967).

Lovaas, O. I., Freitag, G., Gold, V. J., and Kassorla, I. C. "Experimental Studies in Childhood Schizophrenia: Analysis of Self-Destructive Behavior." Journal of Experimental Child Psychology, Vol. 2, (1965), pp. 67-84, (a).

Lovaas, O. I., Freitag, G., Gold, V. J., and Kassorla, I. C. "Recording Apparatus and Procedure for Observation of Behaviors of Children in Free Play Settings." Journal of Experimental Child Psychology, Vol. 2, (1965), pp. 108-120, (b).

Lovaas, O. I., Schaefer, B., and Simmons, J. Q. "Building Social Behavior in Autistic Children by Use of Electric Shock." Journal of Experimental Research in Personality, Vol. 2, (1965), pp. 99-109.

Moustakas, C. E., Sigel, I. E., and Schalock, H. D. "An Objective Method for the Measurement and Analysis of Child-Adult Interaction." Child Development, Vol. 27, (1956), pp. 109-134.

O'Leary, K. D., O'Leary, S. G., and Becker, W. C. "Modification of a Deviant Sibling Interaction Pattern in the Home." Behavior Research and Therapy, Vol. 5, (1967), pp. 113-120.

Ricketts, Agnes, F. "A Study of the Behavior of Young Children in Anger." Studies in Child Welfare, University of Iowa, Vol. 9, (1934), pp. 160-171.

Wahler, R. C., Winkel, G. H., Peterson, R. F., and Morrison, D. C. "Mothers as Behavior Therapists for Their Own Children." Behavior Research and Therapy, Vol. 3, (1965), pp. 113-124.

Wright, Herbert F. "Observational Child Study." P. H. Mussen, editor, Handbook of Research Methods in Child Development, (1960), p. 99.