Reliability of Recording as Affected by Behavioral Consequences

LeRoy J. Switlick

Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses

Part of the Psychoanalysis and Psychotherapy Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/2563
RELIABILITY OF RECORDING
AS AFFECTED BY
BEHAVIORAL CONSEQUENCES

by

LeRoy J. Switlick

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

Western Michigan University
Kalamazoo, Michigan
August 1974
ACKNOWLEDGEMENTS

I wish to express my gratitude to Professors Hermann Peine, Neil Kent, Paul Mountjoy, and Robert Hawkins for their encouragement and constructive criticism throughout this study. Recognition is also extended to James Kirklin for his technical assistance and to the Child Development Center for allowing me to video tape one of their clients. I am also indebted to my wife, Kitty, for her assistance in preparing the video tapes and the manuscript. These acknowledgements in no way detract from the responsibility I assume for the contents of this paper.

LeRoy J. Switlick
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
# TABLE OF CONTENTS

INTRODUCTION ......................................................... 1  

METHOD ............................................................... 16  

Subjects ........................................................ 16  

Apparatus and Setting ............................................ 16  

Procedure ........................................................ 19  

General procedure ............................................ 19  

Training session .............................................. 19  

Sessons I-IV ................................................ 19  

Dependent variables .......................................... 22  

RESULTS ............................................................. 23  

Visual Inspection ................................................ 23  

Reliabilities ................................................ 23  

Intervals-of-occurrence ..................................... 23  

Statistical Analysis ............................................ 32  

Control Group reliability across sessions ................... 32  

Group differences in reliability within sessions .............. 32  

Group differences in intervals-of-occurrence ................ 36  

Sequential effects of the independent variable .............. 37  

DISCUSSION .......................................................... 39  

REFERENCES ......................................................... 46  

APPENDICES .......................................................... 49
# LIST OF TABLES AND FIGURES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Experimental Conditions</td>
<td>20</td>
</tr>
<tr>
<td>II. Analysis of Variance on Reliability Scores</td>
<td>35</td>
</tr>
<tr>
<td>III. Two Factor Analysis of Data From Sessions II and III</td>
<td>38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reproduction of the McLaughlin and Malaby Reliability Data With Vertical Division Lines and Mean Reliabilities Added</td>
<td>11</td>
</tr>
<tr>
<td>2. Mean Reliabilities for Each Group Across Sessions</td>
<td>24</td>
</tr>
<tr>
<td>3. Mean Reliabilities for Each Group Across 10-Min. Tape Segments</td>
<td>26</td>
</tr>
<tr>
<td>4. Group Mean Frequencies of Intervals of Occurrence Across Sessions</td>
<td>28</td>
</tr>
<tr>
<td>5. Group Mean Frequencies of Intervals of Occurrence Across 10-Min. Segments</td>
<td>30</td>
</tr>
<tr>
<td>6. Group Mean Deviations From the Criterion Pen-graph Across 10-Min. Segments</td>
<td>33</td>
</tr>
</tbody>
</table>
INTRODUCTION

Behavior modification, which is the extension of the various learning theories, draws much of its power through the utilization of empirical methodology (Bandura, 1969; Tharp and Wetzel, 1969). The deliberate manipulation and control of relevant variables that are correlated with a particular behavior affords a systematic analysis of those factors which might be responsible for the behavior's occurrence or dimensions (Sidman, 1960; Skinner, 1938; Skinner, 1953). By definition, empirical methodology demands objective observation and data collection. Unless the accurate and objective recording of events surrounding a behavior can be accomplished, little can be concluded about the function of those events.

With this consideration, early investigations of behavioral relationships typically occurred in a controlled laboratory setting. Behaviors under study were operationally defined in such a way as to enable automated recording and to preclude as much human error as possible. Whenever possible, mechanical or electro-mechanical recording devices were utilized. Behaviors were often regarded as "events" that activated the recording device (Skinner, 1938). Those behaviors which did not lend themselves to automated recording were avoided in experimental design.

However, much of current applied behavior analysis deals with behaviors which do not lend themselves to automated recording. Whereas the laboratory affords replication and statistical analysis, the
natural environment often doesn't (Baer, Wolf, and Risley, 1968). Such behaviors as "disruptive-behavior", "self-injurious-behavior" or "face-touching", regardless of how explicitly defined, present special recording problems. Some behaviors, by their complex nature, cannot be recorded by an automated device because of technological limitations. For example, a device which would automatically record "inappropriate-verbalizations" exclusive of the concurrent "decision-making" of a human observer is beyond current technology and economy. Some behaviors such as "disruptive-behavior-in-the-classroom" might be defined to include several component behaviors such as "getting-out-of-seat" and "shouting." Such behaviors are usually referred to as response classes and the multiplicity of topography precludes mechanized observational techniques (Romanczyk, Kent, Diament, and O'Leary, 1973).

The current emphasis on extending behavior modification procedures to the natural environment with multiple stimulus conditions and response components demands data collection procedures other than the ones available to the laboratory. The reliable use of humans to quantify the behavior of other humans is an area of psychological technology long since well developed and very often necessary to applied behavior analysis (Baer, et al., 1968). Human observation as a recording device presents several problems.

Since target behaviors must often be described topographically and involve a certain amount of decision, several controls must be implemented. Assurance must be obtained that the events the observer is recording are the same ones that the experimenter is interested in
and the behavior(s) must be adequately defined to allow replication. The observer must be capable of detecting the behavior. It must be determined that the observer's behavioral definitions and recording behaviors remain constant throughout the recording procedure. Collateral observer behavior must not interfere with the recording behavior. How well the experimenter controls for each of these variables will determine the accuracy and utility of his data.

One method of assessing the reliability of data is to obtain more than one record of the target behavior from more than one observer. These records might then be compared to determine the degree of agreement. Typical of applied behavioral studies is the interval method of recording where the observer records the occurrence or non-occurrence of a behavior within a predetermined time interval. An observer usually records a sequence of such intervals to determine a particular behavior's rate. This method of recording affords the opportunity for another observer to simultaneously and independently record data for purposes of reliability assessment. These additional data are then compared with the primary observer's data, interval by interval, and converted to a percentage of agreement. If the percentage is high, the experimenter has reason to conclude that his data are reliable and proceeds with his interpretations.

However, the validity of this method of reliability assessment has been questioned. A common method of computation is to divide the number of agreements (of occurrence and non-occurrence) by the total number of intervals in the session and to convert to a percentage. A closer analysis of this procedure demonstrates that if the rate of the
target behavior is low, relative to the total number of recording intervals, the percentage of agreement could be high regardless of whether the two observer's ever agree that the behavior occurred during the same interval. Conversely, if the rate of the target behavior is relatively high, a high percentage of agreement could result regardless of whether the two observers are, in fact, recording precisely the same behavior during the same interval. Some researchers have recognized these limitations and have advocated a more stringent computation of observer reliability.

Bijou, Peterson, and Ault (1968) suggest that only those intervals in which either one or both observers record a behavior's occurrence should be used in the computation. Hawkins and Dotson (1972) further suggest that two reliability coefficients be computed—one to indicate the degree of agreement on the behavior's occurrence, and one to indicate the degree of agreement on the behavior's non-occurrence.

Even with more stringent computational procedures, the experimenter is still not assured that a high reliability coefficient indicates accurate data. A high reliability coefficient might be misrepresentative in those studies where only an occasional reliability check is made and where a single reliability coefficient is computed to represent average reliability across all experimental conditions. Thus, for example, in an ABAB design with an inter-observer reliability of 90%, reliability for Condition A might be 80% in each case and reliability for Condition B might be 100% in each case. If the lower reliabilities for Conditions A is the result of a consistent underestimation by the primary observer, an apparent experimental effect could
be nonexistent. Single reliability coefficients to represent overall reliability are, in fact, occasionally reported in recent studies (eg. Bolstad and Johnson, 1972; Christopherson, Arnold, Hill, and Quilitch, 1972; and White, Nielsen, and Johnson, 1972). Other studies provide the range of observer agreement as well as the mean reliability but do not indicate which conditions produced the lower reliabilities (eg. Harris and Sherman, 1973; Nordquist and Wahler, 1973; and Santogrossi, O'Leary, Romanczyk, and Kaufman, 1973).

Even in some studies where separate coefficients are computed for each condition (eg. Kazdin and Klock, 1973; Keilitz, Tucker, and Horner, 1973; and Moore and Bailey, 1973), there is no indication as to whether the disagreement between the two observers was in the same direction across all conditions. It is possible, for example, in an ABAB design with an inter-observer reliability coefficient of 90% for each condition and with an occasional reliability checker recording accurately, to still get biased results (Hawkins and Dotson, 1972). If the primary observer consistently underestimates the behavior's rate in Conditions A by 10% and consistently overestimates the behavior's rate in Conditions B by 10%, there would result a 20% variation in recording behavior even though all conditions yielded the same reliability coefficient of 90%. Thus, recording variability is not necessarily detected by the reliability coefficients per se, and the experimenter has to look at the reliability data and determine whether differences in agreement are randomly distributed or whether they are differentially biased. Although it is likely that most experimenters do, in fact, look closely at the data, the specific sources and param-
eters of inter-observer disagreements are seldom reported.

Observer bias can affect data in several ways. For example, in an ABAB design in which the experimenter predicts that a behavior's rate will increase during Conditions B, there are at least four ways that the data might become biased: (1) the observer might consistently inflate the behavior's rate for all conditions; (2) the observer might consistently deflate the behavior's rate for all conditions; (3) the observer might deflate the baseline rate, inflate the treatment rate, or both; and (4) the observer might inflate the baseline rate, deflate the treatment rate, or both.

In the first two cases, although the data might not be highly accurate, some experimental conclusion might still be made in that a consistent overestimation or underestimation would make relative comparisons between the different conditions still possible. This, of course, would depend upon the degree of error.

In the latter two cases, where observer bias is not constant for all conditions, experimental conclusion might be very difficult. In the case where the observer deflates baseline rate or inflates treatment rate, any experimental effect is exaggerated and the experimenter has little way of determining how much of it is due to treatment effects and how much is due to observer error. If the experimenter is unaware that there has been an observer bias, he might conclude that there was an experimental effect, when in fact there was none. In the case where the observer inflates the baseline rate or deflates the treatment rate, any experimental effect is diminished and the experimenter has little way of determining how much of this was due
to no treatment effect and how much was due to observer bias. Again, if the experimenter is unaware of an observer bias, he may make an invalid conclusion. In this case, however, he may wrongly conclude that there was little or no experimental effect, when in fact there was one. In the former case, the experimenter may wrongly reject a null hypothesis and a Type I error would occur. In the latter case, the experimenter might wrongly accept a null hypothesis and a Type II error would occur (Glass and Stanley, 1970).

It is apparent from the above discussion that coefficients of reliability can be inadequate and insufficient for purposes of assessing the accuracy of recorded data. Bijou, Peterson, Harris, Allen, and Johnson (1969) point out at least four determinants of agreement or observer reliability: (1) adequacy of the observational code; (2) training of the observer; (3) method of calculating reliability coefficients; and (4) frequency of observations over sessions. Hawkins and Dotson (1972) describe three independent sources of error in obtaining accurate and objective data: (1) the observer may be poorly trained, unmotivated, or otherwise incompetent; (2) the definition of the behavior may be vague or incomplete; and (3) the behavior may be difficult to detect because of its subtlety or complexity. Hawkins and Dobes (1973) showed the difficulty in getting reliability without explicit definitions. Azrin, Holz, Ulrich, and Goldiamond demonstrated falsification of data by ambitious or otherwise unmotivated observers. Rosenthal (1966) described the effects of experimenter expectancy on obtained results. Fixen, Phillips, and Wolf (1972) demonstrated that self-reporting and peer-reporting can be very unreliable. Reid (1970)
showed a tendency for observers to adjust their recording behavior to correspond with that of a second observer.

These considerations suggest that the experimenter must ask not merely, was behavior changed?, but also, whose behavior? (Baer, et al., 1968). If humans are used to record a specific behavior, then any change in the data might reflect a change in their observing and recording behavior, rather than in the subject's behavior. Assessment of observer reliability by comparison with another observer does not necessarily indicate the accuracy of the data—it merely demonstrates the degree of agreement between observers. It is possible, even with well trained observers, that both observers concurrently deviate from the behavioral definitions as a function of the experimental conditions.

For example, in a single subject design, where the subject serves as his own control, the stimulus conditions for the observer are not identical for all conditions. During baseline conditions the subject may be behaving at a free operant level without obvious external interaction or manipulation, whereas obvious external interaction in the form of reinforcement or punishment may occur during treatment conditions. The observer might then respond to the reinforcement or punishment as well as to the behavior. This might either interfere with or facilitate the observation and recording task. In either case, if the recording behavior does in fact change as a function of the sequence of conditions, then accuracy of recording will not be constant across conditions and experimental conclusion will be difficult.

There are several reasons that the observing and recording behavior might be different during experimental manipulation (as com-
pared to baseline). There may be increased motivation in the observer in that he might view the manipulation as more crucial than the baseline condition and consequently attend more to his task or be more stringent in behavioral criterion. Consequation during manipulation might also begin to serve as a signal to the observer that the behavior has occurred and thereby acquire discriminative stimulus properties. Consequation could even have reinforcing properties for the observer's behavior as well as the subject's.

The potential for consequating events to become discriminative stimuli is suggested in a study by McLaughlin and Malaby (1972). In a token economy classroom, they investigated procedures to decrease the rate of inappropriate verbalizations. Two manipulations were utilized in an ABAB design: (1) Point Loss, wherein whenever the teacher observed an inappropriate verbalization, the student was immediately requested to remove points from his point chart; and (2) Quiet Behavior Points, wherein if a student engaged in quiet behavior for a 30-min. period, he received bonus points. The sequence of conditions was: Point Loss I, Quiet Behavior Points I, Point Loss II, and Quiet Behavior Points II. Two class members served as independent observers and recorded every occurrence of inappropriate verbalizations.

Apparent results were that while a stable rate of the target behavior occurred during Point Loss I, the rate substantially decreased across Quiet Behavior Points I sessions. Furthermore, when Point Loss (I) was reinstated, inappropriate verbalizations increased across sessions, and when Quiet Behavior Points (II) was reinstated, the target behavior's rate abruptly decreased.
In addition to assessing behavior change procedures, the investigators also attempted to assess the efficacy of using elementary students as observers. Daily reliability coefficients were computed to determine the degree of agreement between the two observers. An inter-observer agreement range of 70 to 100% is cited. Mean agreement for all observations was 89%. As a part of their assessment, these investigators plotted a graph of the daily reliability coefficients. Figure 1 (p. 12) is a reproduction of their graph. Analysis of their graph results in the observation that out of the sixteen Point Loss sessions, reliability was less than 88% only once and mean reliability appears to be approximately 92%. In contrast to this, an overview of the Quiet Behavior Points sessions shows that 12 of the 24 reliability coefficients are less than 88%, with at least six of them less than 80%.

This apparent difference in reliability patterns as a function of the type of experimental condition in effect takes on special significance when it is recalled that in the conditions with the higher reliabilities the teacher immediately consequated the behavior, whereas in the conditions with the lower coefficients, the behavior was not consequated until the end of the session. It is possible that immediate consequation served as a discriminative stimulus for the observers. When consequation was delayed until the end of the session, it could not serve as a discriminative stimulus. The two different schedules of consequation could have resulted in differential observer behavior and the apparent experimental conclusions may have been in reality an artifact.
Figure 1
Reproduction of the McLaughlin and Malaby Reliability Data
With Vertical Division Lines and Mean Reliabilities Added

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The above observations suggest a further analysis of observer behavior as a function of differential stimulus discrimination across conditions. Stimuli which might acquire the properties of a discriminative stimulus are those which indicate a verbal consequence to a target behavior. Such auditory stimuli as verbal praise or reprimand might acquire discriminative properties if the latency between the target behavior and the verbal consequence is of shorter latency than the latency between the target behavior and the observer's recording behavior. (If the latency between the target behavior and the verbal consequence is of longer latency than the latency between the target behavior and the observer's recording behavior, the verbal consequence might acquire the properties of a conditioned reinforcer.) If an auditory stimulus does acquire the properties of a discriminative stimulus for the recording of the target behavior, then the resultant observer data will demonstrate reliability no greater than the correspondence between the auditory stimulus and the behavior's occurrence. If the auditory stimulus occurs each time the target behavior occurs, then the recording data might still be accurate, although not necessarily based on the observation and discrimination of the target behavior.

Whether or not an auditory stimulus demonstrates or acquires discriminative stimulus properties is especially important in those designs where the verbal stimulus (consequence) is presented according to either an intermittent or a multiple schedule.

Auditory stimuli serving as a discriminative stimulus in an intermittent schedule would deflate the recording frequency during
those conditions in which that schedule is in effect. This deflation would not occur during conditions without that schedule such as baseline or reversal. The result would be inconsistent reliability across conditions.

If an auditory stimulus which demonstrates discriminative stimulus properties does not occur each time the target behavior occurs, and furthermore, occurs at times when the target behavior does not occur, the resultant recorded data might bear little relationship to the actual target behavior's rate. This situation might exist in those procedures where the subject is being verbally consequated for two or more similar, but not identical, behaviors both with the same verbal consequence and both on an intermittent schedule. If the observer is to record only one of the behaviors and is unaware that more than one behavior is being consequated, he might use the verbal consequation as a discriminative stimulus and record an occurrence of the target behavior every time the auditory stimulus occurs, regardless of whether it corresponds with the behavior he is recording. Furthermore, with repeated exposure to this condition, the observer might begin to adjust the behavioral definition to include more than one behavior during conditions without the auditory stimulus (such as reversal). This latter possibility is supported in studies by Reid (1970) and Romanczyk, et al. (1971) in which observers demonstrated a shift in definition to correspond with a second observer.

The purpose of this study was to assess the effects of an auditory stimulus on the recording behavior of observers whose task it was to observe and record the frequency and duration of a target behavior.
Three groups of observers recorded the frequency and duration of the target behavior from a set of video tapes. Each group viewed the same tapes and in the same sequence. However, the audio portion of the video tapes differed for each group. Two groups were exposed to auditory stimuli which either represented consequation of the target behavior on a continuous schedule or represented consequation of two similar behaviors on an intermittent schedule. The third group received no auditory stimuli representative of consequation to either behavior.

Several questions were considered:
1. Do auditory stimuli which represent consequation of a behavior demonstrate or acquire discriminative stimulus properties?
2. Do auditory stimuli which resemble (but do not represent) consequation of a behavior demonstrate or acquire discriminative stimulus properties?
3. If auditory stimuli acquire discriminative stimulus properties during a continuous consequation schedule, do these properties generalize to other schedules?
4. If auditory stimuli are initially presented in such a schedule that they do not acquire discriminative stimulus properties, will experience with such schedules prevent or retard the acquisition in subsequent schedules?
5. If auditory stimuli acquire discriminative stimulus properties resulting in altered recording behaviors, do recording behaviors resume a pre-discriminative stimulus level when the auditory stimulus is no longer presented?
METHOD

Subjects

Twenty-one students from an undergraduate psychology course participated in this study for bonus points. All subjects were without prior recording experience and were naive to the purpose of this study. The subjects were randomly assigned to one of three groups and participated in five daily 40-min. sessions. One subject did not complete the study and his partial data were excluded from the results.

Apparatus and Setting

Four 30-min. video tapes were made with a Sony Video-corder (Model AV-3400) equipped with a zoom lens and a directional microphone. Taping sessions occurred at a pre-sheltered workshop for trainable retardates. The subject was a 16-yr.-old mongoloid male who had been observed to emit a high rate of tongue protrusions. Taping occurred while the subject was seated at a table working at various fine motor tasks. The video recorder was positioned approximately ten feet from the subject and at the subject's eye level. The zoom lens was adjusted so that the upper portion of the subject's body (arm pit to top of head) covered the vertical distance of the monitor screen. Of the four 30-min. tapes made, one tape was used for training purposes, and the remaining three were used for experimental manipulation. Two of the experimental tapes (Tapes 2 and 3) were duplicated so that different audio tracks could be superimposed on identical video tapes.
Tape 1 remained intact. The sound track of the duplicated tapes was transferred to a Sony Cassette-corder (Model TC-66).

Assessment of the tapes and experimental sessions occurred in a psychology laboratory which was equipped with a Zenith 24" (diagonal) monochromatic television. Positioned 5½ feet in front of the TV were three student desks which were separated by 4-ft. high cardboard barriers. Four push-button recording devices were constructed to measure 2½"x4"x2". These recording devices were circuited to a pen-recorder which was located in an adjacent room. This room was equipped with an electric fan which masked any noise made by the deflection of the pens on the pen-recorder. One recording device was also circuited to an array of six ¼-inch diameter stimulus lights which were located four inches below the TV screen. The pen-recorder was geared to move the graph paper at a rate of ⅛-inch per 10-sec.

The frequency and duration of tongue-protrusions (Appendix A) were determined by the experimenter for each tape by utilizing the recording apparatus described above. Each tape was viewed on the laboratory TV six times and the resultant pen-graphs were segmented into 180 ten-sec. intervals from common starting points. Corresponding pen-graphs were compared and disagreements as to whether a tongue-protrusion occurred in any particular interval were noted. Intervals of disagreement were then re-monitored and a determination of occurrence or non-occurrence was made. This procedure produced for each tape a criterion pen-graph of 180 successive 10-sec. intervals which indicated the occurrence or non-occurrence of a tongue-protrusion for each interval.
New sound tracks for Tapes 2 and 3 were made so that each tape regained its original sound track but with an added verbal consequation upon every occurrence of a tongue-protrusion. New sound tracks were also made for the duplicates of Tapes 2 and 3. These tapes also regained their original sound tracks but with an added verbal consequation to each tongue-protrusion during one-half (randomly selected) of the intervals in which at least one tongue-protrusion occurred. Further, the sound tracks to the duplicate tapes were made to include a verbal consequation for the similar but not identical behavior of mouth-open (Appendix B) during those intervals in which a tongue-protrusion had not occurred. The verbal consequation for both behaviors was a highly discriminable "No!" and occurred within one second after the outset of the behaviors. This short latency afforded a high probability that consequation would occur before any recording response.

For those tapes in which consequation was to occur following every tongue-protrusion, the experimenter synchronized the tape recorded sound track with the video tape, and by monitoring the video portion was able to simultaneously dub in the original sound track plus the verbal consequations. For those tapes in which consequation was to occur for two different behaviors in different intervals, the experimenter predetermined and noted on the criterion pen-graphs which intervals would include consequations. Then through the synchronization of the video tapes with a stop watch, the experimenter was able to dub in the appropriate consequations by having an assistant cue him as to when to consequate which behavior. As with the other two
Procedure

General procedure.

All subjects monitored and recorded from a video tape once a day for five consecutive days. Due to time limitations, the subjects were run in groups of three. To avoid mutual cueing, the subjects were separated by cardboard barriers and noise from the recording apparatus was masked as described above. The first session was a training session and the four subsequent sessions were for experimental manipulation. The experimental sessions were labeled Session I, II, III, and IV. The video tapes that the subjects recorded from were different for Sessions I thru III. The video tape for Session IV was the same one as Session I. Table I (p. 20) lists the various conditions and sequence for each group of subjects.

Training session.

The subjects were instructed as to the type of behavior to be recorded (tongue-protrusions), advised that they would be collecting data to be used in the assessment of a behavior modification procedure, and shown the training tape to practice on. The specific instructions and procedures are presented in Appendix C.

Sessions I-IV.

The subjects were instructed that they were to record the oc-
### Table I

Experimental Conditions

<table>
<thead>
<tr>
<th>Group</th>
<th>Session I (Tape 1)</th>
<th>Session II (Tape 2)</th>
<th>Session III (Tape 3)</th>
<th>Session IV (Tape 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>no auditory</td>
<td>auditory</td>
<td>auditory</td>
<td>no auditory</td>
</tr>
<tr>
<td></td>
<td>consequation</td>
<td>consequation</td>
<td>consequation</td>
<td>consequation</td>
</tr>
<tr>
<td></td>
<td>one behavior</td>
<td>one behavior</td>
<td>two behaviors</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>no auditory</td>
<td>auditory</td>
<td>auditory</td>
<td>no auditory</td>
</tr>
<tr>
<td></td>
<td>consequation</td>
<td>consequation</td>
<td>consequation</td>
<td>consequation</td>
</tr>
<tr>
<td></td>
<td>two behaviors</td>
<td>two behaviors</td>
<td>one behavior</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>no auditory</td>
<td>no auditory</td>
<td>no auditory</td>
<td>no auditory</td>
</tr>
<tr>
<td></td>
<td>consequation</td>
<td>consequation</td>
<td>consequation</td>
<td>consequation</td>
</tr>
</tbody>
</table>
currences of tongue-protrusions from a thirty-minute video tape. They were reminded to press the recording button on each occurrence of the target behavior and were requested to review the definition prior to the start of the tape. Appendix D contains the specific instructions that were read to the subjects.

During Session I, all three groups recorded from Tape 1. As indicated above, this tape retained its original sound track with no alterations.

During Session II, all groups recorded from Tape 2. Group I heard the auditory stimulus ("No!") immediately following each occurrence of a tongue-protrusion. Group II recorded from the duplicate tape which presented the same auditory stimulus, but only during one-half of the intervals. Additionally, they heard the auditory stimulus immediately following occurrences of mouth open behavior during those intervals that a tongue-protrusion had not occurred. Group III (Control) recorded while hearing only the original sound track. (This was accomplished by playing the video portion of the tape in synchronization with the sound track on the cassette recorder.)

During Session III, all groups recorded from Tape 3. Group I heard the auditory stimulus according to the schedule that Group II did during the previous session, Group II heard the auditory schedule of Group I in the previous session, and Group III again heard only the original sound track.

During Session IV, all three groups again recorded from the tape of Session I (Tape 1) and, as in Session I, heard no superimposed consequences.
Dependent variables.

The measure of recording variability was the reliability of recording of each subject during each session. This was determined by transforming each 30-min. pen-graph into 10-sec. intervals and calculating reliability coefficients by the scored interval method in which the number of agreements (with the standard) were divided by the number of agreements plus disagreements and multiplying times one hundred. Thus each subject provided four reliability coefficients to correspond to each of the four sessions.

Further, a more sensitive measure of moment-to-moment reliability was made by dividing each 30-min. session into three 10-min. segments. Thus, twelve additional reliability coefficients were able to be computed on each subject to assess changes in reliability within sessions.

In addition to the above measures of reliability, the total number of intervals in which at least one response was recorded was determined for each subject for each session and for each 10-min. segment. These frequency counts were compared with each other and with the standard.
Results

Visual Inspection

Reliabilities.

Mean reliabilities for each group were plotted across sessions (Figure 2, p. 25) and across 10-min. segments (Figure 3, p. 27). Although group differences were not apparent in Session I, considerable differences did occur in subsequent sessions. In Sessions II and III, the particular group that heard the auditory stimulus upon each occurrence of the target behavior recorded more reliably than the other two groups. Within group comparisons suggested that the intermittent schedules did not produce reliabilities lower than the baseline session (Session I). Further, Control Group reliability declined subsequent to Session II, whereas the reliabilities of both treatment groups did not.

Intervals-of-occurrence.

An interval-of-occurrence was defined as any 10-sec. tape segment during which at least one tongue protrusion was recorded. The total number of intervals-of-occurrence was tabulated for each subject across sessions and across 10-min. segments. Group means were calculated and are presented in Figures 4 (p. 29) and 5 (p. 31). In Session I, group differences were not apparent. However, beginning with Session II, Control Group appeared to record fewer intervals-of-occurrence than either treatment group. To further explore this
Figure 2

Mean Reliabilities for Each Group Across Sessions
Figure 3
Mean Reliabilities for Each Group
Across 10-Min. Tape Segments
Figure 4

Group Mean Frequencies
Of Intervals of Occurrence Across Sessions

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 5

Group Mean Frequencies
Of Intervals of Occurrence
Across 10-Min. Segments
phenomenon, the actual number of intervals-of-occurrence was obtained
from the criterion pen-graphs. Mean deviations from criterion for
each group were then determined and plotted across 10-min. segments
(Figure 6, p. 34). Although group differences were not noticeable
in Session I, they subsequently became apparent. Whereas Control
Group tended to deviate in the negative direction, deviation scores
of both treatment groups appeared to be distributed around zero.

Statistical Analysis

**Control Group reliability across sessions.**

A one-way analysis of variance was performed on the reliability
scores of Group III (Control) across sessions to investigate differ­
ences in reliability as a function of time or differences in tape
quality. This resulted in significance at the 0.05 level of confi­
dence ($F=5.58$, df=3, 20). Mean reliabilities for respective sessions
were:Session I=80%, -II=84%, -III=75%, and -IV=65%. The obtained
Control Group variance argued against within group statistical analy­
sis of the independent variable.

**Group differences in reliability within sessions.**

A one-way analysis of variance was performed on the reliability
scores of the three groups within each session. The results of these
analyses are summarized in Table II (p. 35). As predicted, no signi­
ficant differences were obtained in Session I ($F=1.26$, df=2, 17, p>.25).
However, Session II demonstrated differences at the 0.05 level of
Figure 6

Group Mean Deviations
From the Criterion Pen-graph
Across 10-Min. Segments
Table II
Analysis of Variance on Reliability Scores

<table>
<thead>
<tr>
<th>Session</th>
<th>F</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1.26</td>
<td>2,17</td>
<td>&gt;.25</td>
</tr>
<tr>
<td>II</td>
<td>3.82</td>
<td>2,17</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>III</td>
<td>5.91</td>
<td>2,17</td>
<td>&lt;.025</td>
</tr>
<tr>
<td>IV</td>
<td>4.02</td>
<td>2,17</td>
<td>&lt;.05</td>
</tr>
</tbody>
</table>
confidence (F=3.82, df=2, 17), Session III at the 0.025 level (F=5.91, df=2, 17), and Session IV at the 0.05 level (F=4.02, df=2, 17).

Further analysis of group differences were then obtained with Sheffe's test for multiple comparisons (Glass and Stanley, 1970). Since Session I differences were nonsignificant, further analysis was not necessary.

In Session II, no significant differences were obtained in Group II vs. Control. However, significant differences were demonstrated in Group II and Control vs. Group I. In this particular session, Group I was the group that heard the auditory stimulus upon each occurrence of the target behavior.

In Session III, significance was not obtained in Group I vs. Control. However, significance was obtained in Group I and Control vs. Group II. In this session, Group II heard the auditory stimulus upon each occurrence of the target behavior.

In Session IV, no differences were noted between Groups I and II. In this session, in contrast to the previous sessions, Groups I and II vs. Control was significant. This session was a return to baseline wherein no group heard the auditory stimulus.

**Group differences in intervals-of-occurrence.**

A one-way analysis of variance was performed on the tabulated intervals-of-occurrence within sessions. In Session I, group differences were nonsignificant (F=0.316, df=2, 17, p>.25). In Session II (F=9.6, df=2, 17) and in Session III (F=10.47, df=2, 17) significance was obtained at the 0.005 level of confidence. In Session IV,
although visual examination suggested differences, statistical analysis produced nonsignificance \((F=2.99, \, df=2, \, 17, \, p>.05)\).

Sheffe's test was then performed on the data from Sessions II and III. In both cases it was observed that Groups I and II vs. Control demonstrated significance, whereas Group I vs. Group II did not.

**Sequential effects of the independent variable.**

As indicated above, Control Group variability across sessions evidenced the presence of confounding variables and limitations in statistical analysis. It was reasoned, however, that a two-way analysis of variance could be performed on the data of Groups I and II in Sessions II and III. By treating Session and Schedule as the two independent factors, information as to sequential effects of the schedules might be obtained. If Session differences were found to be nonsignificant, then any interaction effect would be indicative of a sequential effect.

The above analysis was performed and the results are summarized in Table III (p.38). As predicted, differences between the two schedules were significant \((F=18.5, \, df=1,24, \, p<.001)\). Differences between the two sessions were not significant \((F=3.08, \, df=1,24, \, p>.05)\). Further analysis demonstrated the absence of any significant interaction \((F=1.01, \, df=1,24, \, p>.25)\).
Table III
Two Factor Analysis of Data
From Sessions II and III

<table>
<thead>
<tr>
<th>Source</th>
<th>F</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schedules</td>
<td>18.5</td>
<td>1,24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sessions</td>
<td>3.08</td>
<td>1,24</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>Interaction</td>
<td>1.01</td>
<td>1,24</td>
<td>&gt;.25</td>
</tr>
</tbody>
</table>
DISCUSSION

The purpose of this study was to investigate the effects of differential stimulus conditions on recording behavior. It was hypothesized that stimuli other than the target behavior's occurrence can serve as discriminative stimuli for recording behavior. The specific stimuli that were dealt with were those produced by the consequation of behaviors. If those stimuli demonstrate control over recording behavior, the accuracy of recording could be seriously affected. Experimental conditions such as baseline and reversal would not include those stimuli. Consequently, comparisons between those conditions which include such stimuli and those which do not would be confounded to some degree by differences in reliability.

The particular stimuli that were manipulated in this study were the auditory stimuli made available to the observers. A highly discriminable auditory consequence was presented according to two different schedules. According to one schedule (continuous), there was a perfect correspondence between that stimulus and the target behavior's occurrence. According to the other schedule (variable), there was only a fractional correspondence. Additionally, the latter schedule included the same consequation to a similar, but not identical, behavior.

Two groups recorded the behavior's occurrence with exposure to both schedules but in a counterbalanced order. A control group recorded the target behavior's occurrence without exposure to either schedule.

The measure of stimulus control was the reliability of record-
ing as determined by the comparison of individual data with a criterion. Although this was an indirect measure of stimulus control, the specific manipulations employed and the inclusion of the control group afforded valid inferences.

Systematic relationships were found to occur between the independent (schedule) and dependent (reliability) variables. Between group comparisons indicated that reliability of recording was significantly higher in those conditions where every occurrence of the target behavior was consequted. Further, reliability returned to its former level when the behavior was no longer consequted.

This relationship, by itself, does not indicate the occurrence of stimulus control. There is an alternative explanation that the mere occurrence of discrete auditory stimuli served as a catalyst for increasing accuracy. The data, however, do not support this alternative. In Session II, Group II was exposed to the stimulus according to a variable schedule. In that session, the reliabilities were not significantly different from those in Control Group. If the specific role of the auditory stimulus was that of a catalyst, it would have been reflected by differences between those two groups in Session II.

Session II data provide further information as to the effects of the auditory stimulation. Group I heard the auditory stimulus in perfect correspondence with the target behavior's occurrence. The topography of that stimulus and its temporal relationship to the behavior made it an apparent consequence. Group II, however, heard the stimulus intermittently following two behaviors. This variable schedule made its role as a consequence less apparent. Group I recorded
more reliably than either Group I or Control. Group II's reliability was essentially the same as Control's. These results suggest that any stimulus control the auditory stimulus had for Group I was the result of conditioning that occurred during the session itself. If the auditory stimulus had stimulus control prior to the session, then the reliability of Group II would have been significantly lower than Control Group, rather than on the same level.

The statistical treatment of Session III data demonstrated a significant difference in Group I and Control vs. Group II. In this particular session, Group II received the auditory stimulus in perfect correspondence, whereas, Group I had the variable schedule. These results are somewhat paradoxical. Since Group II had received the stimulus according to a variable schedule in the previous session, one might reason that this group would have learned that the stimulus did not occur in perfect correspondence with the target behavior. Consequently, one might expect this group to progressively attend less to the occurrences of the auditory stimuli. This would then decrease the probability that stimulus control would be demonstrated in subsequent sessions. As the data indicates, however, this group did attend to the auditory stimulus when it occurred in perfect correspondence to the target behavior.

Further, if Group I had, in fact, demonstrated the effects of stimulus control in Session II, these effects might be expected to continue in Session III. Since Group I was exposed to consequating stimuli for two behaviors in Session III, a continuation of stimulus control would have been reflected by low reliability as compared to
Control Group. This did not occur. This suggests that, for both experimental groups, a discrimination as to whether the auditory stimulus corresponded to the occurrence of the target behavior was made early in the sessions. Once this discrimination was made, the observer could react to the stimulus' occurrence accordingly.

This latter possibility gains support from the two factor analysis of variance which was performed on the reliability scores of Groups I and II over Sessions II and III. It was demonstrated that those conditions with auditory consequation on a continuous schedule produced significantly higher reliabilities than those conditions with a variable schedule. However, the combined data of Session II contrasted to the combined data of Session III did not demonstrate differences. These results indicated that any sequential effect of the two schedules would be demonstrated by an interaction effect. No significant interaction was obtained. Since no sequential effect was apparent, one might conclude that the acquisition or non-acquisition of auditory stimulus control occurred within each session, regardless of previous histories of auditory stimulation.

The analysis of the intervals-of-occurrence data lends further support to the potential for auditory stimulus control. In Sessions I and IV, no significant differences were obtained. However, during Sessions II and III, it appeared that Control Group recorded fewer intervals-of-occurrence than both experimental groups during corresponding 10-min. segments. From Figure 6, it can be observed that, whereas Control Group tended to underestimate the frequency of tongue-protrusions, Groups I and II did not markedly deviate from the criterion.
These results would be predicted if the auditory stimulus did, in fact, acquire discriminative stimulus properties for these two groups. It will be recalled that the number of intervals in which auditory stimuli occurred was equal to the actual number of intervals-of-occurrence, regardless of their temporal correspondence. Thus, by merely responding to the auditory stimuli, the observers could record the correct number of intervals regardless of whether they ever attended to the target behavior. Whether this actually occurred cannot, however, be concluded from the available data.

In addition to the above considerations, the data also suggest that the occurrence of the auditory stimuli, although not always facilitating, were sufficient to at least maintain reliability. With the exceptions of those conditions wherein the auditory stimulus were presented on a continuous schedule, reliabilities for both experimental groups maintained a fairly constant level. One consideration is that the occurrence of the stimuli maintained interest and motivation. The presentation of such stimuli gave the appearance that a behavior modification procedure was actually being implemented. Since Control Group had no exposure to those stimuli, their performance might be expected to deteriorate through boredom. In fact, Control Group showed a progressive decline in reliability subsequent to Session II.

Another consideration is that the presentation of auditory stimuli provided additional training for the observers. Specifically, in those conditions where the behavior was consequated upon every occurrence, the observers were afforded external feedback as to their accuracy. Since Control Group experienced no such feedback, the apparent
decline in their performance might be a result of extinction.

In general, there was considerable evidence that observers attend to stimuli other than the target behavior's occurrence while recording. It was further evidenced that these stimuli can affect reliability. However, there are some methodological characteristics which warrant consideration.

Although precautions were made to prevent inter-observer cueing, occasional verbal comments between observers may have affected their attention to the auditory stimuli. It is possible that comments during the variable schedules may have acted to terminate previous stimulus control. This might account for the absence of deteriorated reliabilities during those conditions. It is suggested, however, that comments during the continuous schedules would not confound the evidence for auditory stimulus control. Regardless of whether the auditory stimulus acquired discriminative stimulus properties by way of a casual comment or by way of continued exposure to the stimulus, the fact that its occurrence increased reliability supports the hypothesis of differential reliability as a function of stimulus conditions.

Another methodological consideration is the short length of time that was spent in training. Although the target behavior was considered to be fairly straightforward, the performance of Control Group across sessions suggests that increased training may have produced different results. If initial reliability had been higher for all groups, then any effect of the auditory stimulus may have become obscured by a ceiling effect. However, higher initial reliability might have then provided a more sensitive measure of stimulus control in the
variable schedules. It is possible that the auditory stimulation did
deteriorate reliability during the variable schedules but that these
effects were obscured by fluctuations and decrements in Control Group's
performance.

It is the author's conclusion that this study has demonstrated
the potential for recording behavior to come under the stimulus con­
trol of consequating events. In this particular study the specific
stimuli dealt with were auditory and the behavior being recorded de­
manded a visual discrimination. Further, this study was a laboratory
analogue to recording procedures which more typically occur in the
natural environment. The extent to which the above observations might
generalize to other settings is subject to further research. Such fac­
tors as the topography of the behavior and of the consequence deserve
further consideration. The number of subjects or behaviors being re­
corded might also affect stimulus control potentials. The difficulty
in detecting the target behavior might also be a factor, in that, if a
behavior's consequence is easier to discriminate than the behavior it­
self, the observer might be well tempted to start attending to the
consequence.

It is hoped that this study will stimulate further research.
It certainly lends further reinforcement to the concept of proceeding
with caution in scientific interpretation.
REFERENCES


Hawkins, R.P., and Dotson, V. Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of inter-observer agreement scores in interval recording. Paper pre-
sented at the Third Annual Symposium on Behavior Analysis, Lawrence, Kansas, May 10, 1972.


Tharp, R.G., and Wetzel, R.J. *Behavior Modification in the Natural*
A tongue-protrusion is defined as any outward tongue thrust that results in at least part of the tongue extending clearly beyond the outermost border of the bottom lip. If the tongue is not clearly extending beyond the bottom lip, it is not a tongue-protrusion. When the tongue is withdrawn so that its tip no longer extends beyond the bottom lip, that tongue-protrusion has terminated.

A tongue-protrusion may be of any duration. The duration as well as the occurrence of all tongue-protrusions will be recorded by depressing the recording button for the duration of the response.

If the tongue is protruding, and, if before it is withdrawn the bottom lip becomes shielded by an external event (e.g., head moves off screen, arm in front of mouth, etc.), the response will end two seconds (count 1001...1002) after the mouth is shielded. If the bottom lip again becomes visible before the response ends (less than 2-sec. shielded) and if the tongue continues to protrude, the recording button will continue to be depressed until the tongue is withdrawn.
MOUTH OPEN DEFINED

Mouth-open is defined as any parting of the lips to the extent that the tongue is visible, but not meeting criterion of a tongue-protrusion. Responses meeting criterion of a tongue-protrusion shall be considered as an occurrence of a tongue-protrusion and not as an occurrence of mouth-open.
APPENDIX C
TRAINING PROCEDURE
AND INSTRUCTIONS

The experimenter read the instructions to the subjects and provided them with a copy of the behavioral definition of tongue-protrusion.

"We are evaluating a method of decreasing tongue-protrusions in a 16-yr.-old retarded boy. As in any evaluation of behavior modification procedures, it is necessary that we get an accurate measure of the behavior's frequency. We are using video tapes of the procedure with hopes of attaining high accuracy. We are also having several people record the behavior because the combined data will have higher accuracy.

"Your task will be to record the frequency of tongue-protrusions by depressing the recording button every time you see a tongue-protrusion occur. This first session is a training session for you to familiarize yourselves with the behavioral definition and with the recording apparatus. This session, and each session hereafter, will last approximately 40 minutes. There will be five sessions and you will receive bonus points for your participation as explained earlier.

"Here is a description of the behavior you will be recording. Please read it carefully. Please feel free to ask any questions." Experimenter gave subjects a copy of definition and answered questions.

"When you are recording, be sure to record independently and not attend to what the other is doing. Just be sure that you press the recording button every time you see a tongue-protrusion occur. 51
"Remember, press the recording button immediately. Do not wait for the response to end. Press the button for the entire duration of each tongue-protrusion.

"Now we will begin training. You will begin observing and recording a 30-min. tape. We have already figured out when the subject makes a tongue-protrusion so that I can tell you how well you are doing. Please notice the array of lights beneath the TV. I will record with you and every time I press the recording button, those lights will light up. This will give you the opportunity to learn how to record according to the definition.

"Now begin recording and press the recording button immediately when you decide that the subject is protruding his tongue. Remember to press for the entire duration of any tongue-protrusion. Remember that if the subject's mouth is shielded during a tongue-protrusion for more than two seconds it is possible that two responses are to be recorded even though it might appear that only one has occurred."

The experimenter began the tape and followed the above procedures. At the end of the tape, he instructed them to return the next day.
APPENDIX D
INSTRUCTIONS: SESSIONS I-IV

Session I Only

"From now on, I will not be able to give you feedback on how well you are doing because we have not studied these tapes. We are going to rely on your recording. We do not have any specific expectations of the procedures and do not know whether the frequency of the tongue-protrusions will change."

Sessions I-IV

"Please read and review the behavioral definition. In two minutes you will begin recording." Subjects then reviewed definition.

"You will now begin recording."

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.