78-1445

WILDMAN, Beth Goldstein, 1951-
THE EFFECT OF BIAS INSTRUCTIONS ON THE JUDGMENTS AND INTEROBSERVER AGREEMENTS OF NAIVE BEHAVIOR OBSERVERS.

The University of North Carolina at Greensboro, Ph.D., 1977
Psychology, clinical

Xerox University Microfilms, Ann Arbor, Michigan 48106

THE EFFECT OF BIAS INSTRUCTIONS ON THE JUDGMENTS AND INTEROBSERVER AGREEMENTS OF NAIVE BEHAVIOR OBSERVERS

by

Beth Goldstein Wildman

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
1977

Approved by

Marilyn T. Erickson
Dissertation Adviser

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Dissertation Adviser Marilyn T. Erickson

Committee Members Rosemery O. Nelson

P. Scott Lawrence

William A. Powers

[illegible]

April 22, 1977
Date of Acceptance by Committee

with standard instructions. Accuracy decreased when subjects were shifted from standard to lenient instructions. No changes in accuracy were found for subjects shifted from standard to strict instructions. Subjects were readily separated into liberal, average, and conservative groups. The number of exact agreements between liberal subjects who received strict instructions and conservative subjects who received lenient instructions was not influenced by shifting from standard to experimental instructions. However, the frequency with which these groups recorded "True" responses became more similar. These results were obtained across slide conditions.

The results of the present research suggested that brief instructions can significantly influence the rating patterns of observers. In addition, individual differences in the rating tendencies of observers were found; that is, subjects were readily divided into three bias groups. These findings, in conjunction with the topographies of changes in the recording of behaviors as a function of instructions, and the lack of differences among the slide conditions support the use of signal detection research as an analogue to behavioral observation.

ACKNOWLEDGMENTS

The author wishes to gratefully acknowledge the support and encouragement of her chairperson, Dr. Marilyn T. Erickson, who aided in all aspects of this research and guaranteed the continuation of the research during the author's absence. Dr. Rosemery O. Nelson, Dr. Richard L. Shull, and Dr. P. Scott Lawrence deserve thanks for their aid in the development and clarification of these ideas. The author sincerely appreciates the time and effort that Dr. William A. Powers devoted to aiding her in the selection of appropriate statistical analyses.

Special thanks are extended to Michelle Lange who spent many arduous hours posing for the slides. Michelle, her brother Scott, and her parents recruited other children and interrupted daily schedules in order to complete the slides.

The author wishes to acknowledge Mary Beth Kelly Monroe, Kenneth Asher, and Mimi Aberg for their help in running subjects, returning equipment, and scheduling subjects. Hal Wildman merits thanks for his help in formulating the study and performing the computer analyses. Finally, the author extends deepest appreciation to Hilda Goldstein for financial and emotional support of this research.

TABLE OF CONTENTS

LIST OF TABLES

Table		Page

Table		Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Behavioral scientists have been interested in increasing their understanding of the variables affecting observational data. Concern with improving the reliability of these data has led to numerous investigations of systematic observation methods. One aspect of observational procedure which has begun to receive research and speculative attention has been the area of observer training. Since 1970, investigators have become increasingly interested in identifying variables that might influence the effectiveness of observer training procedures for improving the extent to which observers agree on their rating of behavior.

Although empirical findings concerning variables relevant to observer training procedures have only begun to appear in the literature, investigators using systematic observation for data collection have become increasingly aware of the need for organized and specific training experiences for their observers. Suggestions for effective training procedures, based on empirical findings and intuition, have recently been made available (DeMaster, Reid, & Twentyman, 1977; Johnson & Bolstad, 1973; O'Leary & Kent, 1973; Romanczyk, Kent, Diament, & O'Leary, 1973; Wildman & Erickson, 1977; Wildman, Erickson, & Kent, 1975).

As early as 1933, Thomas, Loomis, and Arrington reported that observers made fewer errors as they gained experience with observation instruments and procedures. These authors also noted that when the rate of disagreement between observers was initially high, repeated use of the observation instrument resulted in more improvement than when observers began with low rates of disagreement. However, until recently there has been a complete lack of research directed toward identifying factors that could enhance the effect of experience alone and maximize the level of observer agreement attainable with training.

DeMaster et al. (1977) supplied one group of observers with information concerning their agreement with their rating partner, as well as with a criterion protocol; a second group of observers was informed of their agreement with their rating partner. Agreement between observers under these conditions was higher than when observers were given rating experience alone. Those observers who received the more complete feedback, that is, reports of their agreement with the standard and their partner's protocols, achieved higher levels of observer agreement at the completion of training than those observers who were given only information concerning how well they concurred with their partner.

Wildman, Erickson, and Kent (1975) extended the findings of DeMaster et al. (1977) by attempting to ascertain whether a specific aspect of feedback during training, consistency, affected the records collected by observers. They found that the agreement percentages obtained by both groups, that is, observers trained by the more consistent standard of a single trainer and those observers trained by the less consistent standard, themselves, were similar. However, Wildman et al. (1975) reported that the observers who were exposed to the more consistent standard were less variable in their ratings of behavior than the self-training group. That is, the variability of the mean number of behaviors recorded per interval was significantly lower for the one-trainer group than for the self-training group.

The findings of Wildman et al. (1975) in conjunction with the results of a study by Romanczyk et al. (1973) form a potent argument for the importance of consistent training standards. Romanczyk et al. (1973) found that observers may adopt different definitions, or interpretations, of specific behaviors as a function of who they believed their reliability assessor (standard) to be. The assessors deliberately scored four of the nine behavioral categories differently from each other. Observers varied their recording of these categories in a manner that complied with the idiosyncratic definitions of the specified reliability assessor.

The findings of these studies have suggested to users of observation procedures that the training of observers is probably a critical factor affecting their data. Recently, investigators who have relied on naturalistic observation for data collection have become more rigorous with respect to the training of their observers. Johnson and Bolstad (1973) described the impressive training procedures which have become standard practice at the Oregon Research Institute. Although many investigators lack training facilities comparable to those detailed by Johnson and Bolstad (1973) in which long periods of time are directed to training observers with daily sessions which include discussion of the code and rating of precoded interactions, an increasing number of researchers have begun to attend to how their observers are trained. In addition, brief descriptions of the observer-training procedures used in a given study have begun to be published in the method sections of articles.

The previously discussed research has clearly increased the interest of those involved with systematic observation, either as a data collection device or as a target for research, in observer-training procedures and research related to these procedures. The issues to which methodology researchers, such as DeMaster et al. (1977), Romanczyk et al. (1973), and Wildman et al. (1975), have addressed their studies have been concerned with identifying types of experiences during training which affect observers' recording of behavior. Observer agreement and accuracy have been used to measure changes in the recording behavior of observers. Observer agreement indicates the degree to which two observers agree with each other, whereas accuracy reflects the degree to which observers agree with a predefined standard.

The independent variables which have been investigated were selected on the basis of intuition or experience.

Although this line of research has been successful in highlighting certain aspects of training procedures which appear to be important, investigators continue to experience difficulty in teaching observers to become proficient enough at their task to achieve and maintain adequate levels of observer agreement (usually 75% to 85% is considered acceptable). Some articles report that data were collected with agreement percentages below the 75% level (e.g., O'Leary, Kent, & Kanowitz, 1975). Some articles which report levels of observer agreement over 75% calculated agreement using liberal procedures (e.g., agreements on occurrences and non-occurrences of behavior). Had more conservative definitions of agreement been used, that is, agreement only on occurrences, observer agreement would have been lower.

It would seem that a more interesting issue for investigation would be the identification of critical components of training experiences that affect observers' behavior. Since current training techniques involve much time and may not be effective in getting observers to agree as much as many investigators would consider desirable, a useful avenue of investigation may be the generation of new, more efficient, training procedures rather than the analysis and improvement of current training procedures. An attempt to achieve an understanding of the fundamental processes involved in observer-training procedures necessitates the formulation of some hypotheses concerning the changes that occur, as a function of training, with respect to responses observers make to relevant stimuli.

A useful basis for these speculations might be the results of research dealing with the perception of discrete stimuli, such as studies concerned with signal detection theory and human vigilance. Studies of human vigilance have concentrated on identifying variables which may influence how well observers detect potent stimuli over sustained spans of time, usually at least one hour in length. A typical observation situation would appear to be analogous to a vigilance task, with the exception of the increased complexity of an observation setting. Signal detection theory can be viewed as an important precursor to a vigilance study since signal detection experiments have the potential for assessing the strength of a stimulus and can, therefore, distinguish between strong and weak signals, or stimuli.

In a typical signal detection experiment (McNicol, 1972), subjects must indicate whether each stimulus presented was either a signal (e.g., pure tone) with noise (e.g., white noise), or noise alone. According to signal detection theory, there are two factors contributing to an observer's responses to stimuli, sensitivity (perceptual ability) and bias (criterion for the classification of a stimulus as a signal). Although an individual's sensitivity tends to remain fairly stable, bias can be varied by giving the observer a new set of instructions concerning the criterion he should apply to his classification of signals or nonsignals. When this change in bias occurs, an observer's responses to the stimuli will be different than they were before his criterion was shifted. A change in bias results in an observer's selecting a different criterion for classifying a stimulus as either a signal or noise. If an observer applies a more liberal criterion, then more stimuli will be classified as signals and fewer will be labeled noise. Conversely, an observer's adoption of a more stringent (or conservative) criterion implies that fewer stimuli will be scored as signals, and consequently more stimuli will be recorded as noise. An investigator can manipulate an observer's bias by instructing him to adopt a different criterion, either a more stringent or a more liberal one.

If one assumes that systematic naturalistic observation can be viewed as a complex signal detection task, then studying observer bias with respect to signal detection theory may provide a useful model for the training of observers. Observers in a naturalistic setting may respond to standard instructions (operational definitions) with individual response tendencies. That is, some observers would tend to record the occurrence of certain behaviors more often than the occurrence of other behaviors, while other observers might record selected categories less frequently than some observers do (Arrington, 1932). Since the goal of training is to obtain agreement between observers, one might hypothesize that observer bias may be shifted during the training experience in such a manner that observers who initially adopt stringent criteria may become more liberal with training, while observers who begin with liberal criteria may become more conservative in their judgments. With the use of the signal detection paradigm, one might further speculate that the manipulation of bias is fundamental to the training of observers, and if a trainer explicitly offered new biases to observers by instructing them to rate differently (e.g., leniently or strictly), they would necessarily change their classification of responses in the desired direction. Observers who approach the rating task with different criteria will probably have relatively poor levels of observer agreement. If observer bias is shifted with instructions, observer agreement is likely to increase.

In the present study, the experimenter attempted to manipulate behavioral recordings directly, by asking observers to change their ratings of a specific behavior. Naive subjects were selected as observers because of their similarity to persons beginning training as behavioral observers. Observers were given specific instructions concerning how behavioral ratings should change with respect to a criterion. Observers were classified on the basis of their initial responses to stimuli and the effect of instructions was viewed in terms of the interaction between the type of instructions and the initial ratings of observers in order to assess whether subjects who began with an extreme rating tendency (e.g., liberal or conservative) could be influenced to shift bidirectionally. Subjects who initially rate at the extremes may not readily shift further in that direction, yet may easily shift in the opposite direction.

In order to permit a test of the applicability of signal detection findings to the complex observation setting involved in naturalistic research, subjects were given different sets of instructions (biases) on which to base their recording of whether or not a given response occurred. In addition, the potential for expanding the use of the signal detection paradigm from the laboratory to the more complex environment of the classroom was assessed by subjecting different groups of observers to the same classification task under three different slide conditions, ranging from a laboratory to a classroom setting. If the signal detection paradigm is limited to relatively simple environments, the effects of instructions would be expected to attenuate as the slides become more complex.

These predictions of change in behavioral ratings as a function of instructions to observers may appear contrary to the findings of observation studies designed to assess the effect of observer bias on observer recordings. These studies (Kent, O'Leary, Diament, & Dietz, 1974; Siegel, Dragovich, & Marholin, 1976; Skindrud, 1972), with the exception of Kass and O'Leary (1970), failed to find significant changes in the ratings of observers as a function of statements to observers that the behavior of target children would change. However, although the previous research and the present study were concerned with similar independent variables, that is, statements to observers, and similar dependent variables, i.e., behavioral recordings, the resemblance between the present research and the past research is limited to these general similarities. Contrary to the current study, the bias research (Kent et al., 1974; Siegel et al., 1976; Skindrud, 1972) has attempted to manipulate behavioral ratings in an indirect manner; that is, the experimenter informed observers that the child's behavior would change, but did not ask the observers to vary how they rated the behavior _per se_. In addition, in the earlier research, observers were given general expectations of behavior change, rather than specific instructions for change. The results of the Romanczyk et al. (1975) study, in which observer recordings varied as a function of reliability assessor, suggest that behavioral observers can change their rating behavior under certain conditions. Finally, the bias research has been oriented toward examining changes that may have occurred in the ratings of a group of experienced, rather than naive, observers.

In summary, the present study investigated whether explicit instructions (lenient or strict) to change the classification of behaviors could influence the rating behavior and interobserver agreement of inexperienced observers.

The interaction between these instructions and the initial response tendencies of observers was examined. In addition, three slide conditions representing a continuum from the simplicity of the laboratory to the complexity of the classroom were evaluated for their effect on observer ratings and interobserver agreement.

CHAPTER II

METHOD

Subjects

The 144 subjects were students enrolled in various undergraduate social science courses at the University of North Carolina at Greensboro. For some, participation in an experiment was part of their course requirement. The others participated on a voluntary basis. The procedure of this study was reviewed by a committee of faculty and graduate students who were responsible for guaranteeing that human subjects are treated in an ethical manner. All subjects were informed about the purpose of the research after they completed the experiment.

Experimental Design

Independent variables. Subjects were assigned to one of three slide conditions and one of two experimental instruction conditions. The slides were of a ten-year-old girl with her head turned at different angles along the horizontal axis. One-third of the subjects (N=48) were assigned to the laboratory condition, in which the child was seated behind a desk, with a white wall behind her. In the classroom without sound condition, the subjects (N=48) viewed slides of the girl seated behind a desk; however, other children were standing next to and behind her. The third group of subjects (N=48) viewed the classroom with sound slides, which were identical to the classroom without sound slides but, in addition, were accompanied by a taped recording of classroom sounds.

Within each of the slide conditions, subjects were assigned to one of two experimental instruction conditions, lenient instructions and strict instructions, with 24 subjects in each. All subjects rated the slides with standard instructions before they were asked to rate with experimental instructions.

After the data collection was completed, the 24 subjects in each of the six groups (3 slide conditions x 2 instruction conditions) were blocked into three groups of subject bias: liberal, average, and conservative. The eight subjects with the least number of "True" responses were classified as conservative, and the eight subjects with the most number of "True" responses were labeled as liberal. The middle eight subjects constituted the average group.

Dependent variables. The dependent measures included the number of times each subject recorded the occurrence of an orienting response (number of "True" responses), the number of times each subject correctly recorded whether or not an orienting response had occurred (number of accurate responses), and observer agreement (the extent to which observers agreed on their classification of responses). For some analyses, differences between the number of "True" (or accurate) responses recorded under experimental instructions and the number of

"True" (or accurate) responses recorded under standard instructions were used. Classification of subject bias was dependent on subjects' responses under standard instructions. Difference scores were used to avoid violating the requirement of independence of variables in an analysis of variance.

Observer agreement was calculated three different ways. One method was to divide the number of agreements on occurrence ("True" responses) by the number of agreements on occurrence plus the number of disagreements. A probability-based formula (Yelton, Wildman, & Erickson, 1977) was also used to calculate agreement:

$$\sum_{Z=A}^{Y} \frac{Y!}{Z!(Y-Z)!} \times \frac{X!}{(X-Z)!} \times \frac{(N-X)!}{((N-X)-(Y-Z))!} \times \frac{(N-Y)!}{N!}$$

where $\underline{A}$ = the number of agreements on occurrence obtained

$\underline{N}$ = the number of intervals

$\underline{X}$ = the number of occurrences recorded by Observer 1

and $\underline{Y}$ = the number of occurrences recorded by Observer 2.

The third method for calculating observer agreement was to divide the number of occurrences recorded by the observer who recorded the smaller number of occurrences by the number of occurrences recorded by the observer who recorded the larger number of occurrences.

Experimental Materials

Three sets of 35-millimeter color slides of a ten-year-old girl with her head turned were made. The slides depicted the girl with her head turned at different angles along the horizontal axis ranging from 15 to 72 degrees in 3-degree increments. In order to position the girl's head appropriately for each angle of orientation, marks were placed on a wall for her to fixate her eyes upon. Adults positioned in the room helped to insure accurate head orientation.

Each angle of orientation was reproduced eight times, resulting in a total of 160 slides which were presented in the same random order for each set of slides. In addition, ten randomly selected poses were placed at the beginning of each set of slides (practice slides), resulting in a total of 170 slides in each set. Slide Set 1 (laboratory condition) depicted the child with her head turned against a white background (see Pose 1, Appendix A). Slide Set 2 (classroom without sound condition) was identical to Slide Set 1 except that the child was seated in a simulated classroom setting with other children adjacent to and behind her (see Pose 2, Appendix A). Both Slide Sets 1 and 2 were not accompanied by sound. Slide Set 3 (classroom with sound condition) was identical to Slide Set 2; however, classroom sounds, recorded on magnetic tape, accompanied the presentation of the slides.

In order to guarantee that the child's head position was constant across slide sets, the only slides actually taken of the child were one picture against a white background (Slide Set 1) for each of the 20 different angles of head orientation. The background was added in a photography laboratory by having technicians superimpose a slide of the child on a slide of a classroom setting (Slide Sets 2 and 3) Eight copies were made of each slide in order to form a set of 160 slides. A cassette recording of classroom sounds was made and accompanied the presentation of Slide Set 3.

Procedure

Subjects were randomly assigned to one of three slide conditions; each slide set was viewed by 48 subjects. Each subject viewed a slide set under two different instruction sets: (1) standard instructions and (2) one of two experimental instruction conditions. Assignment to a particular experimental instruction set was made on a random basis. Each group of observers was shown the slide set for the first time after hearing a tape with the standard instructions.

> An orienting response is defined as the child turning her head at least 45 degrees, that is, 45 degrees or more from straight ahead. Blacken in "True", "T", on your answer sheet if the child's head is turned 45 degrees or more. That is, blacken "T" if the orienting response has occurred. Blacken "F" if an orienting response has not occurred, that is, the child's head is not turned at least 45 degrees. Here is a picture of a 45 degree turn. (Subjects were shown a slide of the child posed at a 45 degree angle for 10 seconds.) Remember, blacken "T" if the child's head is turned 45 degrees or more than 45 degrees. More than 45 degrees means that her head is turned toward the back. Blacken "F" if her head is turned less than 45 degrees. Less than 45 degrees means that her head is turned toward the front.

Subjects were then shown the ten practice slides in order to familiarize them with their task and the procedure for marking their IBM score sheets (see Appendix B).

Subjects' responses to these slides were not included in the data analyses. Following these practice slides, the entire slide set was viewed. Slides appeared for 1.5 seconds, followed by 3.5 seconds for recording. When the subjects completed observing the set, the experimenter collected the answer sheets and gave subjects new answer sheets. After identifying information was written on the sheets, a tape was played of either the lenient or strict instructions.

Lenient instructions:

> It is important that all occurrences of the orienting response be recorded. This time, I would like you to record "True" if you even think that an orienting response has occurred. Although this may seem similar to what you were doing before, this time I am asking you to apply a more liberal definition so that occurrences of the orienting response will not be overlooked. That is, I am asking you to rate more leniently than you did before. However, this does not imply that all responses should be rated "True".

Strict instructions:

> It is important that only real occurrences of the orienting response be recorded. This time, I would like you to record "True" only when you are sure that an orienting response has occurred. Although this may seem similar to what you were doing before, this time I am asking you to apply a more strict definition so that nonoccurrences of the orienting response will not be scored as occurrences. That is, I am asking you to rate more conservatively than you did before. However, this does not imply that all responses should be rated "False".

A maximum of six subjects observed the slides at any one time, until all 144 subjects were run. Seats were assigned so that the angle and distance between the observer and slide screen were controlled.

CHAPTER III

RESULTS

The means, standard deviations, and ranges for each of the two dependent variables, number of times each subject responded "True" to the slides and the number of slides each subject correctly classified (accuracy), as well as the correlation between these variables appear in Table 1. The means and correlations are presented for each of the two instruction conditions (lenient and strict) as well as by slide condition (laboratory, classroom without sound, and classroom with sound) and by subject bias (conservative, average, and liberal). Means for subject performance under standard and experimental (lenient or strict) instructions are also presented.

In order to ascertain the effect of the three slide conditions and the three levels of subject bias on the two dependent variables, two multivariate analyses of covariance, (one for lenient instructions and one for strict instructions), and their concomitant univariate analyses were performed on the data (see Figure 1). F ratios for all multivariate analyses were calculated using Roy's maximum root criterion (Harris, 1975).

The multivariate analyses of covariance contained two fixed between-group factors, slide condition and subject

Table 1
Means, Ranges, and Standard Deviations for
Number of "True" Responses and Accuracy and
Correlation between Number of "True" Responses and Accuracy

Laboratory condition	"True" Responses			Accuracy			
	Mean	Range	Standard deviation	Mean	Range	Standard deviation	r
Lenient instruction condition							
Standard instructions							
Liberal subjects	106.75	100-124	7.65	124.50	76-138	20.99	-.40
Average subjects	93.75	86-100	5.54	142.25	134-147	3.28	-.92**
Conservative subjects	67.37	53-78	10.47	133.38	71-148	25.87	.69*
Lenient instructions							
Liberal subjects	124.75	104-145	14.82	109.00	75-136	20.08	-.57
Average subjects	97.63	71-115	13.57	136.86	125-147	7.82	-.69+
Conservative subjects	99.63	66-146	23.16	112.38	23-146	44.30	-.26
Strict instruction condition							
Standard instructions							
Liberal subjects	101.88	96-112	5.59	137.13	124-144	6.73	-.99***
Average subjects	90.37	81-95	5.50	142.13	128-150	7.22	.17
Conservative subjects	74.13	53-80	9.54	144.38	127-152	7.85	.82*
Strict instructions							
Liberal subjects	92.00	54-112	19.39	131.25	94-145	16.32	.54
Average subjects	72.25	38-91	21.50	137.75	113-144	14.42	.98***
Conservative subjects	60.75	42-75	12.23	136.75	116-149	11.59	.96**

Table 1 (continued)

	"True" responses			Accuracy			
	Mean	Range	Standard deviation	Mean	Range	Standard deviation	r
Classroom without sound condition							
Lenient instruction condition							
Standard instructions							
Liberal subjects	102.50	97-109	4.11	132.50	99-143	14.03	-.78*
Average subjects	88.75	79-95	5.95	144.50	138-150	4.50	-.06
Conservative subjects	67.63	54-79	9.30	129.88	98-149	17.32	.54
Lenient instructions							
Liberal subjects	111.75	104-122	6.04	128.00	118-136	6.00	.99***
Average subjects	106.50	89-124	13.05	131.25	114-143	12.20	-.96**
Conservative subjects	75.50	46-108	20.72	121.75	70-146	29.76	.19
Strict instruction condition							
Standard instructions							
Liberal subjects	111.50	97-147	15.05	126.25	93-143	16.18	-.94**
Average subjects	85.75	78-95	6.41	144.00	136-151	5.01	.13
Conservative subjects	62.50	54-77	7.43	111.75	35-145	35.85	.34
Strict instructions							
Liberal subjects	106.50	95-131	12.78	127.00	91-144	18.98	-.82*
Average subjects	74.13	53-100	16.02	142.86	133-151	6.53	.58
Conservative subjects	58.50	20-112	28.25	118.75	88-141	18.42	.50
Classroom with sound condition							
Lenient instruction condition							
Standard instructions							
Liberal subjects	107.38	102-113	4.17	125.88	93-137	14.33	-.63+
Average subjects	97.25	90-101	3.69	137.75	123-143	6.18	.27
Conservative subjects	68.63	36-87	16.10	126.88	59-151	29.29	.50

Table 1 (continued)

	"True" responses			Accuracy			
	Mean	Range	Standard deviation	Mean	Range	Standard deviation	r
Classroom with sound condition							
Lenient instruction condition							
Lenient instructions							
Liberal subjects	118.75	106-133	9.10	119.75	107-134	8.84	-.97**
Average subjects	110.50	84-151	19.12	125.50	77-138	20.36	-.94**
Conservative subjects	89.75	45-133	34.08	121.25	76-147	23.96	.05
Strict instruction condition							
Standard instructions							
Liberal subjects	107.25	101-115	5.01	127.25	108-138	11.31	-.77*
Average subjects	98.63	96-101	2.00	138.38	132-144	4.17	-.39
Conservative subjects	73.63	46-86	12.55	130.38	80-150	24.44	.11
Strict instructions							
Liberal subjects	94.50	71-121	17.05	133.50	113-146	13.83	-.44
Average subjects	85.75	62-108	15.23	142.00	132-149	5.07	-.27
Conservative subjects	63.25	41-80	14.18	129.75	85-149	20.74	.09

+ $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Slide Set	Subject "Bias"	Instruction Set			
		Standard		Experimental (Lenient or Strict)	
		Number of "True" responses	Number of Accurate responses	Number of "True" responses	Number of Accurate responses
Laboratory	Liberal				
	Average				
	Conservative				
Classroom without Sound	Liberal				
	Average				
	Conservative				
Classroom with Sound	Liberal				
	Average				
	Conservative				

Figure 1. Experimental design for multivariate analyses of covariance and analyses of covariance.

bias, with three levels each. The dependent measures for these analyses were the number of "True" responses and the number of accurate responses. Since the classification of the levels of subject bias was determined by the number of "True" responses made by each subject under standard instructions, an analysis of covariance with the number of "True" responses recorded under standard instructions as the covariate was necessary. The number of "True" responses and the number of accurate responses recorded under standard instructions were used as the covariate for difference scores obtained by subtracting the number of "True" responses and the number of correct responses recorded under experimental instructions from the number of "True" responses and the number of correct responses recorded under standard instructions.

The results of the multivariate analysis of covariance for lenient instructions appear in Table 1C (all statistical tables appear in Appendix C). The only significant finding of this analysis was the slide condition x bias interaction, F (4, 61) = 2.60, $p < .05$. This same interaction, slide condition x bias, was the only significant finding of the analysis of covariance on the change scores for the number of "True" responses recorded, F (4, 62) = 2.68, $p < .05$, shown in Table 2C. A Newman-Keuls post-hoc analysis indicated that with laboratory slides, conservative subjects changed more in their recording of "True" responses than did average subjects ($p < .01$). The difference between the magnitude of change found for liberal and average subjects approached significance ($p < .10$), with liberal subjects changing more than average subjects did. Conservative subjects changed significantly more in their rating of "True" under the laboratory slide condition than they did in the classroom without sound condition ($p < .01$). See Figure 2 for the graph of this interaction. No significant findings were obtained from the analysis of covariance for the change in accuracy scores (see Table 3C).

A multivariate analysis of variance containing three between-group factors, slide condition, subject bias, and experimental instructions (lenient and strict) was performed in order to examine differences between subjects who received lenient and strict instructions (see Figure 3). The dependent measures in this analysis and the accompanying univariate analyses were the number of "True" responses and the number of accurate responses recorded under experimental instructions. Tables 4C, 5C, and 6C contain the multivariate analysis of covariance and the accompanying analyses of covariance for strict instructions. As can be seen, these analyses did not result in any significant findings. The main effect of bias approached significance, $F(2, 61) = 2.75$, $p < .10$, in the multivariate analysis of covariance.

The results of the multivariate analysis of variance comparing responses under lenient and strict instructions appear in Table 7C. The main effects of instructions,

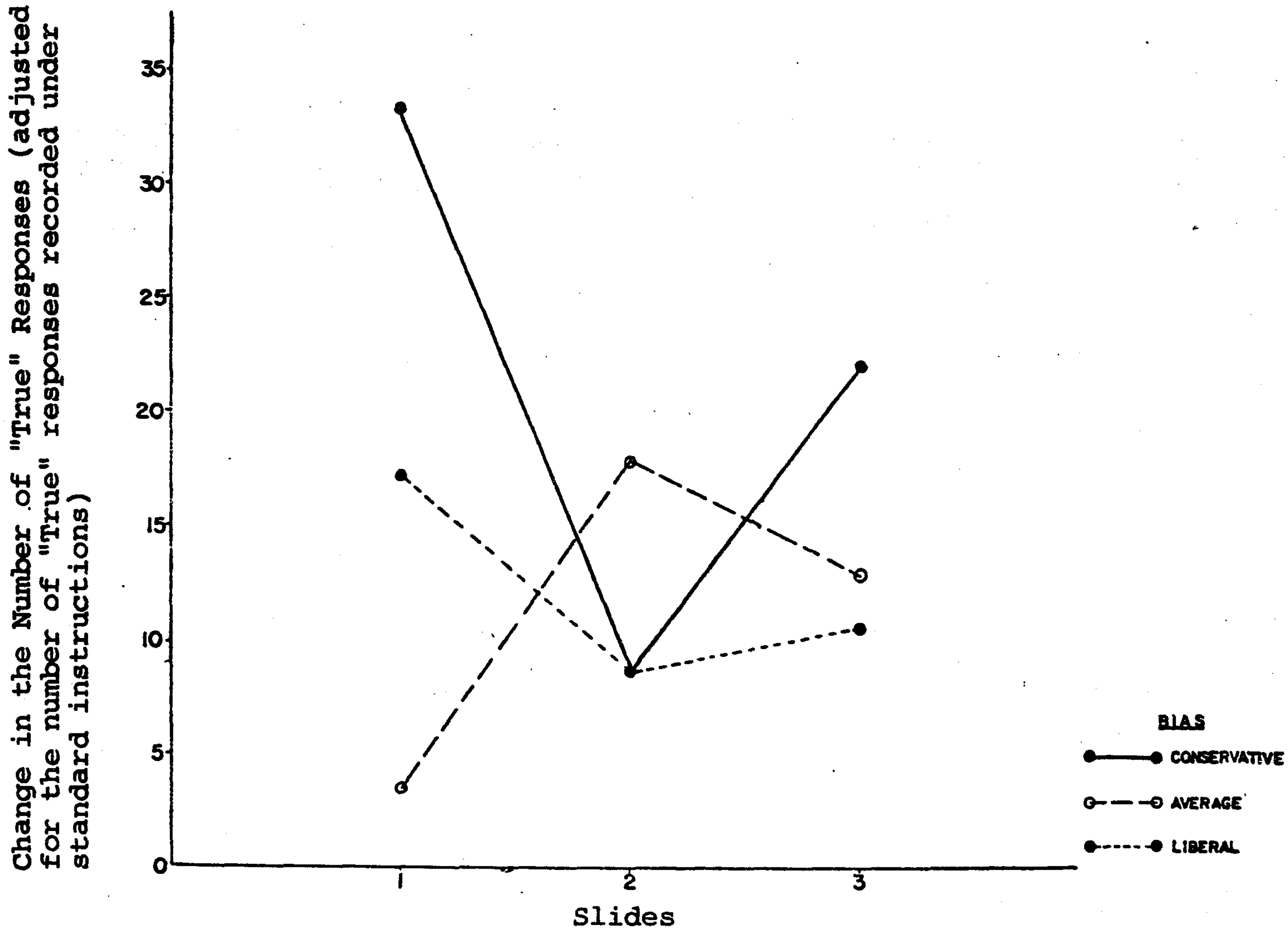

Figure 2. Interaction between slide condition and subject bias for subjects receiving lenient instructions.

Slide Set	Instruction Set	Subject "Bias"	
Laboratory	Lenient	Liberal	
		Average	
		Conservative	
	Strict	Liberal	
		Average	
		Conservative	
Classroom without Sound	Lenient	Liberal	
		Average	
		Conservative	
	Strict	Liberal	
		Average	
		Conservative	
Classroom with Sound	Lenient	Liberal	
		Average	
		Conservative	
	Strict	Liberal	
		Average	
		Conservative	

Figure 3. Experimental design for multivariate analysis of variance and analyses of variance.

F (1, 126) = 75.55, p <.001, and bias, F (2, 126) = 39.96, p <.001, were significant. A significant interaction was found between slide condition and instructions, F (2, 126) = 3.11, p <.05. In addition, the slide condition x instructions x bias interaction approached significance, F (4, 126) = 2.14, p <.10.

In the analysis of variance on the number of "True" responses, the main effect of instructions was found to be significant, F (1, 126) = 67.48, p <.001, with more "True" responses recorded under lenient instructions than under strict instructions. In addition, the main effect of bias was also found to be significant, F (2, 126) = 39.59, p <.001 (see Table 8C). A Newman-Keuls post-hoc analysis comparing the number of "True" responses recorded by subjects with different biases was performed. The results indicated that average subjects recorded significantly more "True" responses than did conservative subjects (p <.001), and liberal subjects recorded more "True" responses than did both average subjects (p <.001) and conservative subjects (p <.001).

The analysis of variance on accuracy scores, which appears in Table 9C, indicated that instructions had a significant effect on the accuracy of subjects' recordings, F (1, 126) = 10.74, p <.001, with strict instructions resulting in more accurate protocols than lenient instructions. In addition, subject bias was also found to affect the accuracy of recording, F (2, 126) = 6.32, p <.01. A Newman-Keuls *post*-*hoc* analysis indicated that average subjects were more accurate than both conservative subjects ($p < .05$) and liberal subjects ($p < .05$).

In order to evaluate the influence of the experimental instructions (lenient or strict) on observers' recordings of the orienting response, six multivariate analyses of variance, with one within-subject and one between-subject factor, were performed using the number of "True" responses recorded and accuracy of recording as dependent measures. The within-subject factor was instruction set, standard or experimental. The slide condition, with three levels, was the between-subject factor. Of the six multivariate analyses, one was performed for each of the three levels of subject bias under each of the two instruction conditions (lenient or strict). Univariate analyses were performed on each of the dependent variables to accompany each of the multivariate analyses.

Tables 10C, 11C, and 12C contain the multivariate analyses of variance for lenient instructions. Subjects of conservative, average, and liberal biases experienced significant shifts from standard to lenient instructions, F (1, 21) = 20.93, $p < .001$, F (1, 21) = 16.40, $p < .001$, F (1, 21) = 45.89, $p < .001$, respectively. The results of the univariate analyses indicated that the number of "True" responses was significantly less under standard instructions than it was under lenient instructions for each of the bias groups, F (1, 21) = 20.92, $p < .001$, F (1, 21) = 13.97, $p < .01$,

$\underline{F}$ (1, 21) = 32.42, $\underline{p} < .001$, for conservative, average, and liberal subjects respectively (see Tables 13C, 14C, and 15C).

According to the results of the univariate analyses, which appear in Tables 16C, 17C, and 18C, accuracy scores decreased for all bias groups as subjects switched from standard to lenient instructions, $\underline{F}$ (1, 21) = 5.74, $\underline{p} < .05$, $\underline{F}$ (1, 21) = 16.24, $\underline{p} < .001$, $\underline{F}$ (1, 21) = 6.99, $\underline{p} < .05$, for conservative, average, and liberal subjects respectively. In addition, the multivariate analysis of variance indicated that slide condition significantly influenced the ratings of liberal subjects, $\underline{F}$ (2, 21) = 4.31, $\underline{p} < .05$. The univariate analysis of variance indicated that the number of "True" responses recorded by subjects with liberal bias was significantly affected by slide condition, $\underline{F}$ (2, 21) = 3.78, $\underline{p} < .05$. A Newman Keuls <u>post-hoc</u> analysis indicated that liberal subjects who viewed laboratory slides recorded significantly more "True" responses than did subjects who viewed the classroom without sound slides ($\underline{p} < .05$).

Tables 19C, 20C, and 21C present the results of the multivariate analyses for subjects who received strict instructions. The ratings of conservative, average, and liberal subjects changed significantly from standard to strict instructions, $\underline{F}$ (1, 21) = 21.62, $\underline{p} < .001$, $\underline{F}$ (1, 21) = 37.76, $\underline{p} < .001$, $\underline{F}$ (1, 21) = 9.35, $\underline{p} < .01$, respectively. The univariate analyses for the number of "True" responses demonstrated that shifting from standard to strict instructions significantly decreased the number of times subjects responded "True" to the slides, $F\ (1, 21) = 5.09$, $p < .05$, $F\ (1, 21) = 19.48$, $p < .001$, $F\ (1, 21) = 9.09$, $p < .01$, for conservative, average, and liberal subjects respectively (see Tables 22C, 23C and 24C). No significant differences for instructions were found in the univariate analyses of variance based upon accuracy scores, which appear in Tables 25C, 26C, and 27C. Slide condition significantly affected recording of conservative and average subjects, $F\ (2, 21) = 4.67$, $p < .05$, $F\ (2, 21) = 6.09$, $p < .01$, respectively.

A univariate analysis of variance indicated that the accuracy scores of conservative subjects were significantly affected by slide condition, $F\ (2, 21) = 3.82$, $p < .05$. A Newman-Keuls *post-hoc* analysis indicated that conservative subjects recorded more accurately in the laboratory condition than in the classroom without sound condition ($p < .05$). For average subjects, the effect of slide condition on subjects' recordings of "True" responses approached significance, $F\ (2, 21) = 3.31$, $p < .10$. No significant differences between slide conditions were found when a Newman-Keuls *post-hoc* analysis was performed.

In order to assess whether experimental instructions influenced observer agreement, an analysis of variance was performed comparing the agreement percentages of randomly assigned pairs of liberal and conservative subjects. Subject pairs were composed of liberal subjects who received standard and strict instructions and conservative subjects who received standard and lenient instructions. These pairs, representing the extremes of the subject distribution, were expected to be maximally different with standard instructions (i.e., poor agreement percentages) and therefore, most likely to increase in agreement with experimental instructions. This analysis included one between group factor, slide condition, and one within subject factor, instruction set, standard and experimental. Arcsin transformations were performed on the observer agreements obtained with each formula before the analyses of variance were done. As can be seen in Table 28C, no significant findings were obtained from the analysis of variance when the agreements divided by agreements plus disagreements formula was used. The analysis of variance in which probability-based agreement scores were used yielded no significant results (see Table 29C). When observer agreement was calculated the third way, by dividing the smaller number of occurrences by the larger number of occurrences, observer agreement under experimental instructions was found to be significantly higher than observer agreement under standard instructions, $\underline{F}$ (1, 21) = 7.34, $\underline{p} < .05$ (see Table 30C). Table 31C contains mean transformed and untransformed agreement percentages using each of the three formulas.

Finally, to permit a signal detection interpretation of the data, the probability of recording "True" and accurate responses for each of the slides was plotted for both liberal and conservative subjects for each of the three slide conditions. Curves for liberal subjects include their number of "True" responses and number of correct responses under both standard and strict instructions. Data for conservative subjects were plotted for their performance under standard and lenient instructions. These graphs appear in Figures 4, 5, 6, 7, 8, and 9. The graphs appear to be very similar across slide conditions, but differ between liberal and conservative subjects. The probability curves for liberal subjects under each of the three slide conditions are nearly identical. The sets of curves for conservative subjects also closely resemble each other. However, the sets of curves for liberal subjects differ markedly from the sets of curves for the conservative subjects. Whereas liberal subjects fell below the 50 percent level (4 "True" responses) in their recording of "True" at approximately Slides 12 to 14 under both standard and strict instructions, conservative subjects fell below the 50 percent level at approximately Slide 9 with standard instructions and fell below this level at higher slide numbers with lenient instructions. Also, liberal subjects tended to shift more abruptly than did conservative subjects from recording primarily "True" to recording primarily "False" responses. Accuracy decreased rapidly and markedly for liberals, and less rapidly and to a lesser magnitude for conservatives.

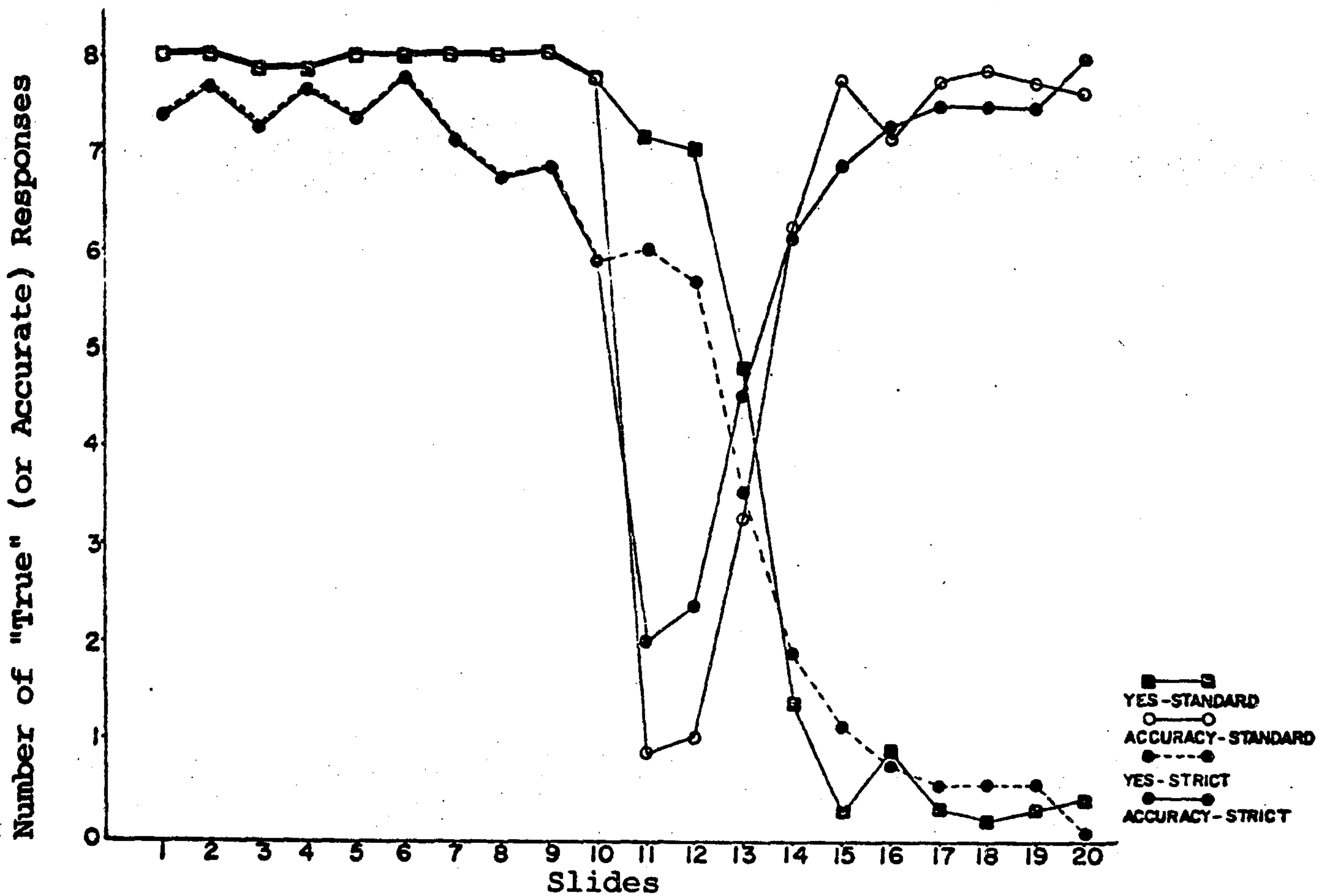

Figure 4. Laboratory condition. Liberal subjects with standard and strict instructions.

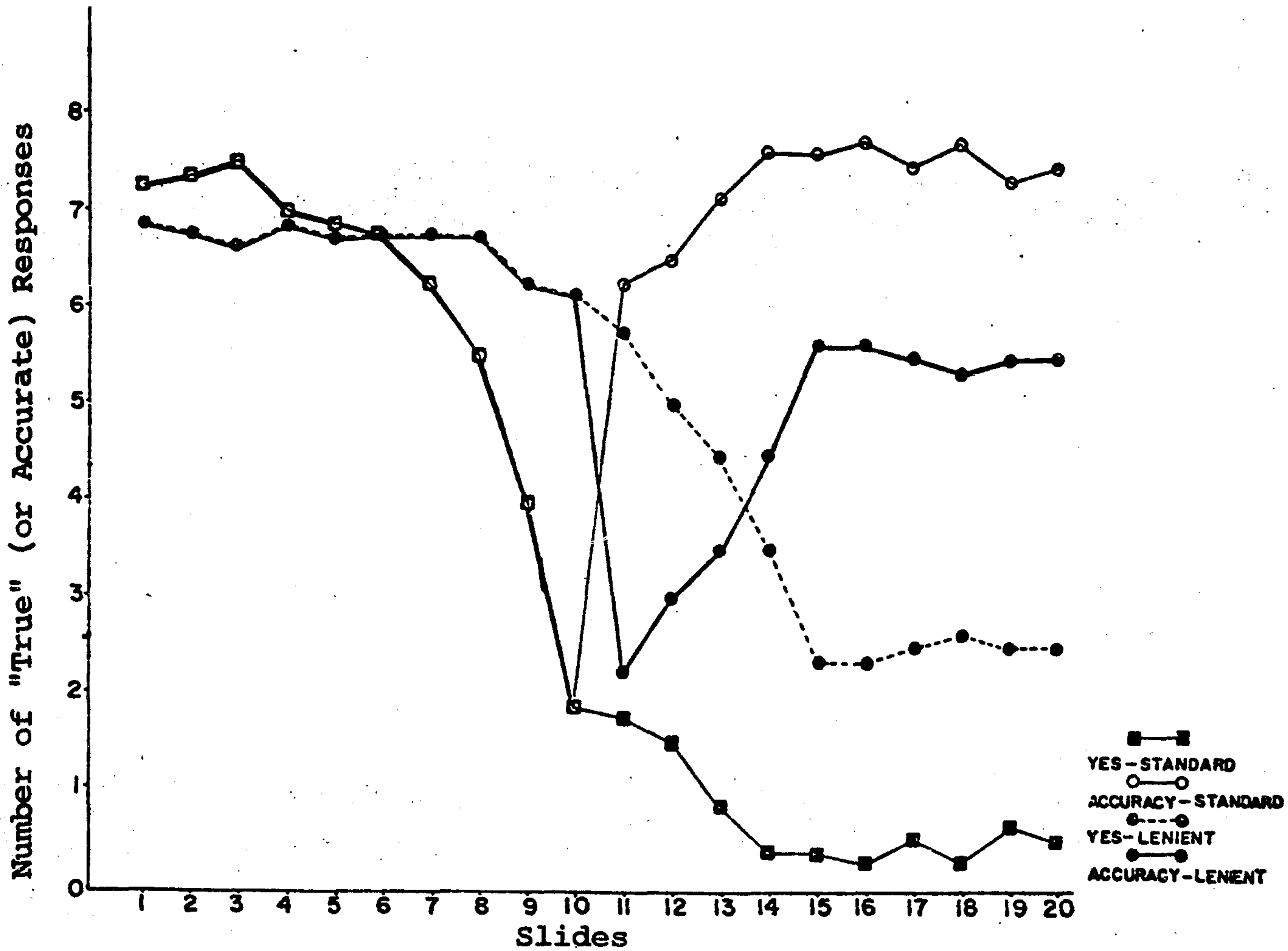

Figure 5. Laboratory Condition. Conservative subjects with standard and lenient instructions.

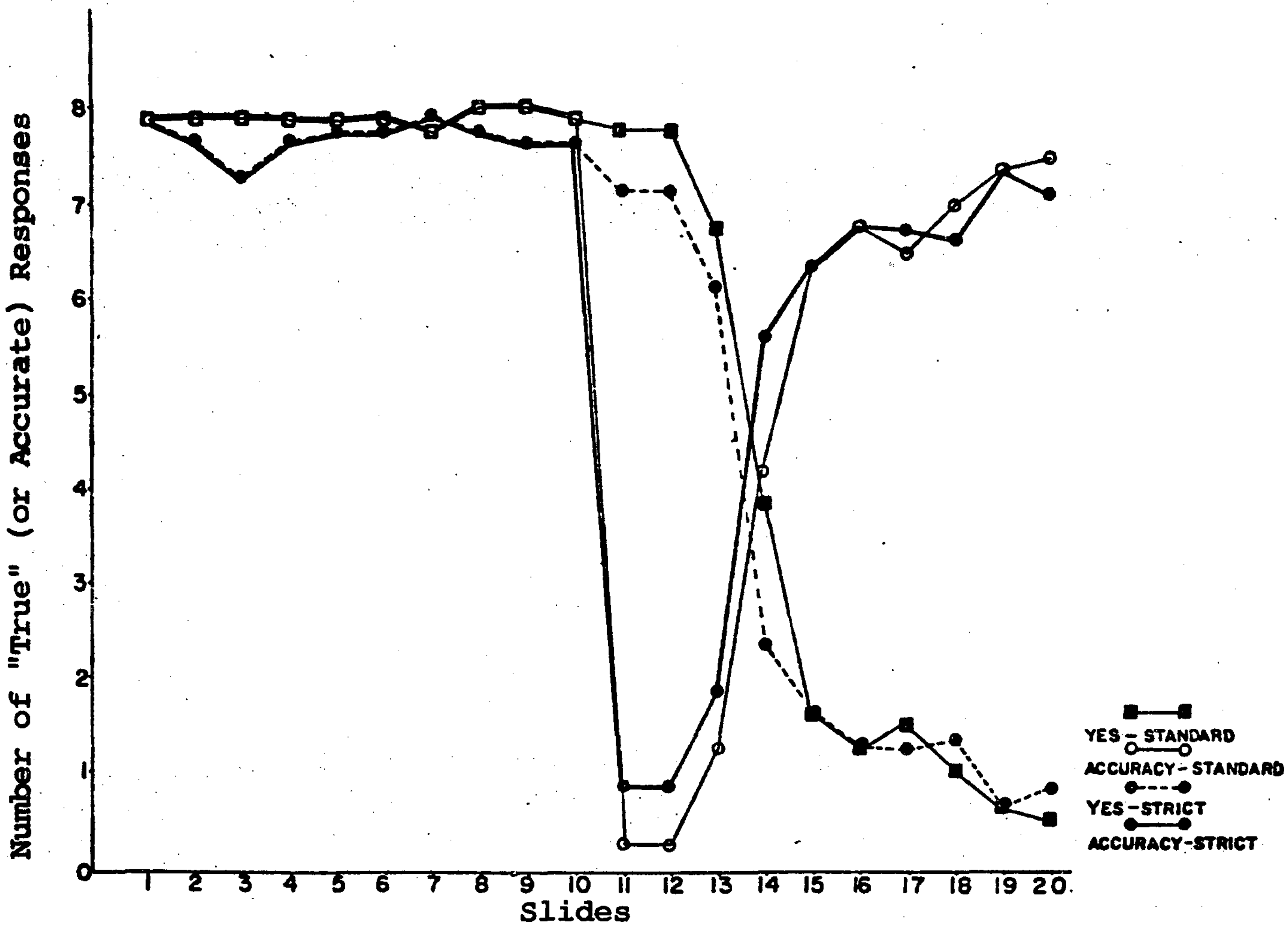

Figure 6. Classroom without sound condition. Liberal subjects with standard and strict instructions.

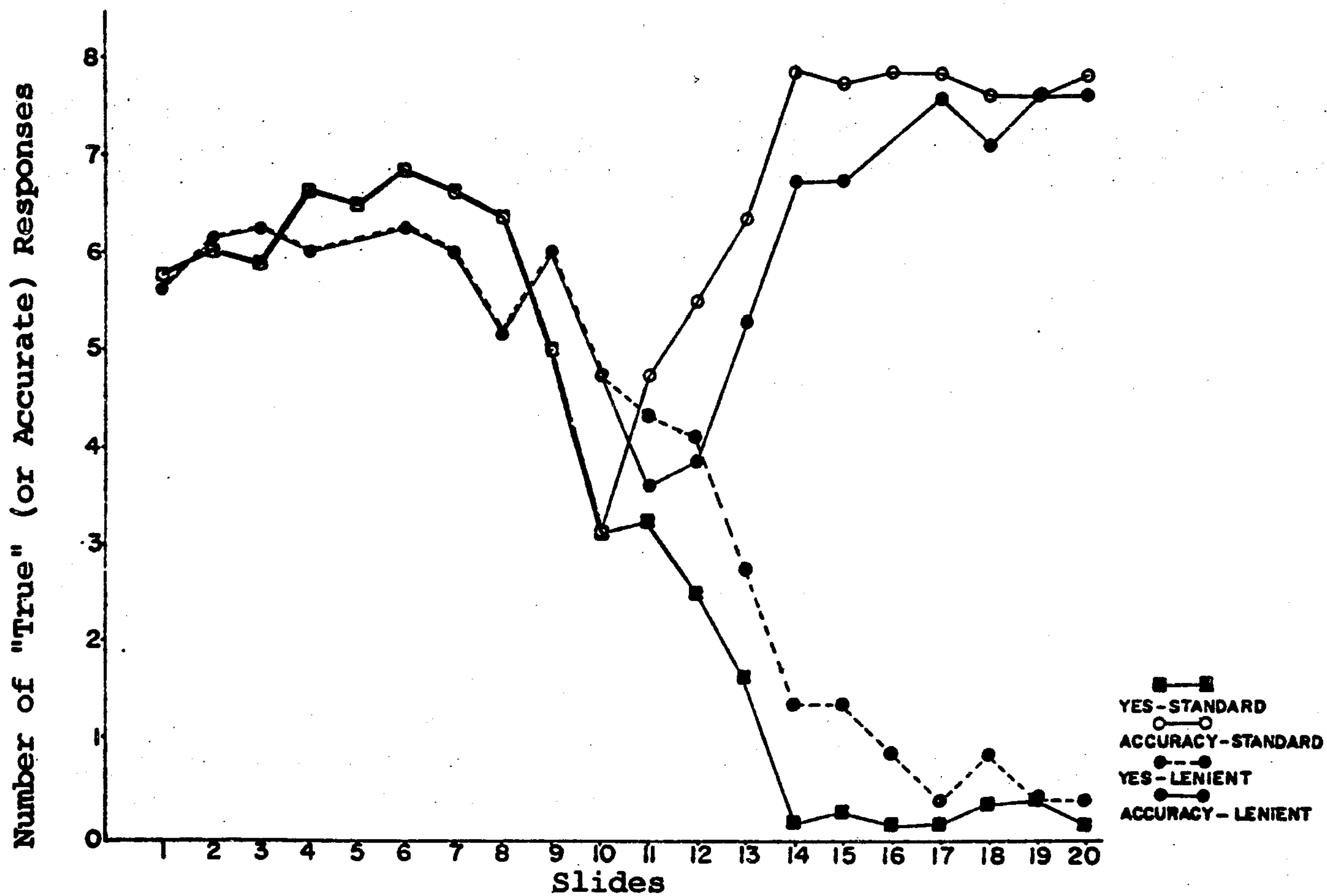

Figure 7. Classroom without sound condition. Conservative subjects with standard and lenient instructions.

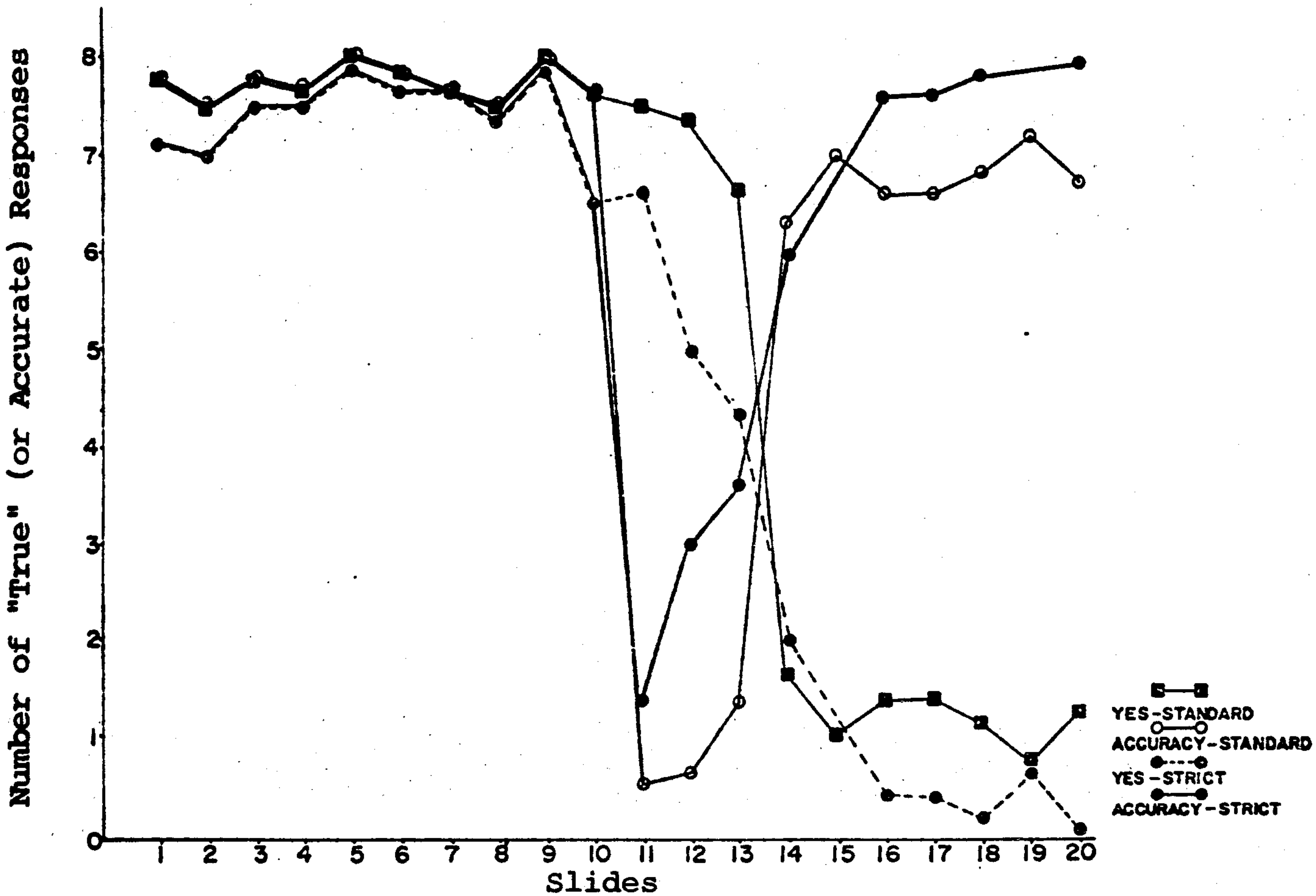

Figure 8. Classroom with sound. Liberal subjects with standard and strict instructions.

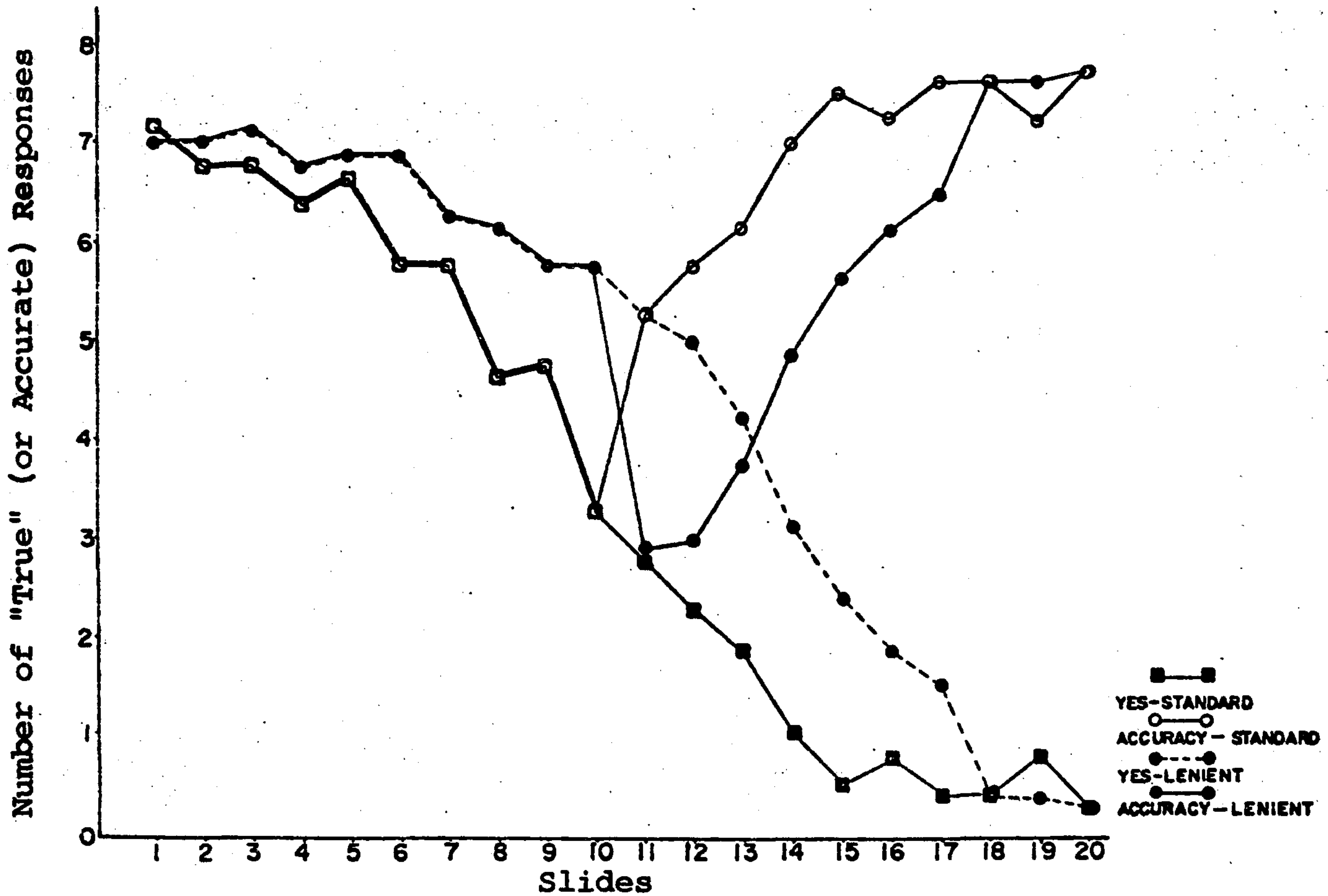

Figure 9. Classroom with sound. Conservative subjects with standard and lenient instructions.

Figures 10 and 11 are the probability curves of two subjects, one liberal and one conservative, from the laboratory slide condition. These graphs, which are representative of the individual data, resemble the group curves.

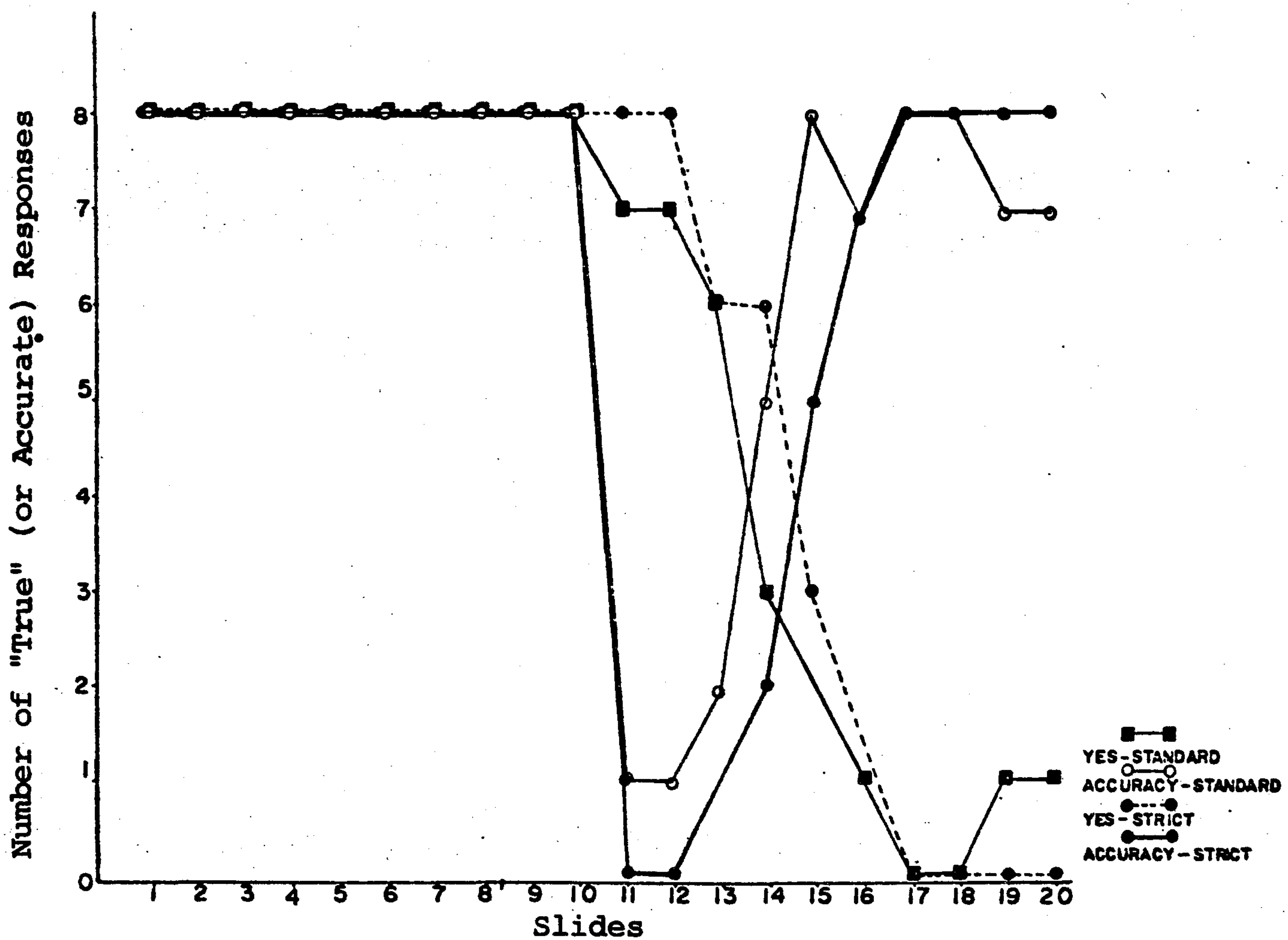

Figure 10. Laboratory condition. Liberal subject with standard and strict instructions.

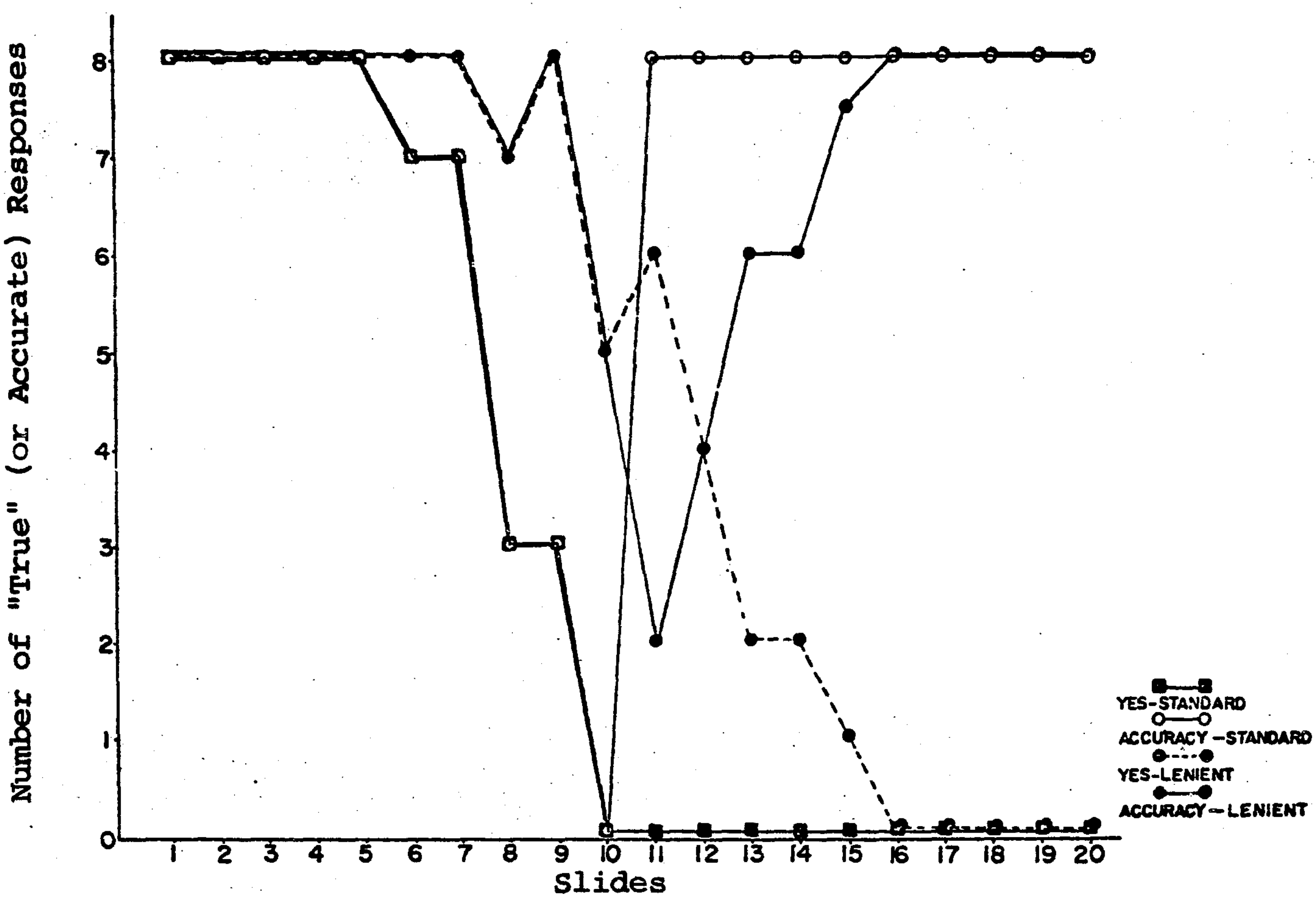

Figure 11. Laboratory condition. Conservative subject with standard and lenient instructions.

CHAPTER IV

DISCUSSION

The findings of the present research are relevant to two aspects of observational research: the influencing, or biasing, of behavioral recordings, and the application of signal detection theory to behavior observation research and training.

Two different aspects of observer bias were explored. First, subjects were grouped on the basis of their response biases, or rating tendencies. The data verified the initial assumption that potential observers do not rate similarly when they begin training. Rather, each observer approaches the observation task from a different starting point, or rating tendency. However, contrary to expectation, subjects did not distribute themselves normally about the hypothetical mean of 80 "True" responses (the orienting response actually occurred 80 times). Instead, most subjects began with liberal response tendencies, such that both liberal and average subjects recorded "True" more than 80 times. The conservative subjects averaged less than 80 "True" responses under standard instructions.

The most obvious explanation for this shift of the distribution is that the slides contained inaccuracies. Careful review of the slides revealed no deviations from consistent increments in the degree of head orientation across the series. However, other aspects of the stimuli may have contributed to the unexpected distribution. On certain slides, the girl's head appeared to be slightly rotated in the vertical plane, resulting in an added cue for identifying certain slides. Aberrations in the color of the slides, as a result of the processing that was necessary, may also have supplied confounding cues.

A second possible explanation for the apparent liberal tendencies of the observers is that the original, or standard instructions were not neutral but, instead, suggested to subjects that they should rate in a lenient manner. However, careful reading of the instructions failed to reveal any implicit or explicit directions to record leniently.

A third hypothesis is that the sample selected did not accurately reflect the population from which it was drawn. This issue can be addressed only after the experiment has been replicated.

The lack of three distinct groups which conform to liberal, average, and conservative standards, relative to the theoretical mean of 80, precludes clear interpretations of some of the findings that will be discussed later. However, although three groups relative to the theoretical mean were not obtained, three distinct groups of subjects who were liberal, average, and conservative relative to one another were obtainable. These groups remained identifiable despite the experimental manipulations. That is, even under lenient and strict instructions, liberal, average, and conservative subjects rated significantly differently from each other.

The second and more common formulation of bias in the context of behavioral observation concerns itself with the ability of an investigator to systematically influence the frequency with which observers record certain behaviors. With respect to this aspect of bias, the present findings clearly document the feasibility of changing recorded frequency in observational records by instructing observers to record differently, that is, to apply a different criterion for judging the occurrence of a behavior.

These results are inconsistent with the literature on observer bias. With the exception of Kass and O'Leary (1970), research has consistently failed to demonstrate that instructions to observers could influence ratings of objectively defined behaviors (Kent et al., 1974; Siegel et al., 1976; Skindrud, 1972). Procedural differences between the current research and the bias research, such as the use of explicit directions to naive observers in the present study could account for the discrepancy between the findings.

However, the present results are consistent with the literature concerned with the training of observers. Although observer training has rarely been conceptualized as a biasing procedure, the goals and procedures of training are directed toward influencing the rating patterns of observers as a means of teaching them to agree with a standard and/or another observer. Instructions to observers during observer training are typically more subtle than those in the present study; yet, they are often successful in producing changes in the rating behavior of observers (e.g., Romanczyk et al., 1973; Wildman et al., 1975).

The principal finding of the present study, relevant to observer training, is the apparent vulnerability of the rating behavior of human observers to a brief manipulation. Subjects significantly changed their classification of a clearly defined overt behavior merely by being asked to change. Observers with initially liberal, average, and conservative response tendencies shifted in the prescribed directions. Examination of the ratings of individual subjects indicated that the majority of subjects changed their ratings in accordance with the instructions, suggesting that the differences that were found were based upon shifts by most observers, rather than by large shifts in only a few observers. These results were obtainable across slide conditions, which represented a continuum from the relative simplicity of the laboratory to the complexity of the classroom with sound.

Since significant changes in the recorded frequency of responses occurred, changes in observer agreement scores were also expected. However, analyses based upon the most frequently used formula (agreements on occurrences divided by agreements on occurrences plus disagreements) failed to demonstrate significant changes in observer agreement data. In addition, no significant changes in agreement scores based upon the probability-based formula (Yelton et al., 1977) were found.

However, when agreement percentages were calculated by dividing the number of occurrences recorded by the observer who recorded the fewest occurrences by the number of occurrences recorded by the observer who recorded more occurrences, significant increases in observer agreement were obtained when observers were shifted from standard to experimental instructions. Conservative subjects recorded more "True" responses and liberal subjects recorded fewer "True" responses with experimental instructions than each had recorded with standard instructions.

These differences in the results as a function of the formula used were unexpected. The agreements over agreements plus disagreements formula does not take into consideration rate of recording (Yelton et al., 1977) and would not be expected to reflect changes unless the proportion of agreements to agreements plus disagreements changed. Although the probability-based formula is responsive to rate of recording, its relationship to other formulas and changes in data is unknown.

Only the formula which defined agreement by frequency of recording was sensitive to the changes in instructions. The insensitivity of the other two formulas may be accounted for on the basis of their requirement that agreement occurs at the same point in time. The lack of increases in agreement using these two formulas may be related to the fact that liberals tended to change most on the slides closest to the 45 degree head turn (Slides 8-12), while conservative subjects tended to shift across the entire range of slides.

These differences suggest limitations in the information supplied by each of the observer agreement formulas. Observer agreement has been used as a measure of the consistency with which observers record behavior. However, in the present research observers significantly changed the frequency with which they recorded behaviors; yet, observer agreement scores, based on exact agreement formulas, did not reflect these changes. These findings suggest that observer agreement is not an adequate measure of the consistency with which observers record.

The data relevant to the aspects of bias discussed earlier, and the accuracy data, as well as the results with respect to changes in observer agreement appear to conform readily to a signal detection analysis. The fact that similar results were obtained across the three slide conditions suggests that the present findings are likely to be applicable to the complex observation environment of the classroom, where much observational research occurs. Consistent differences between slide conditions were not obtained, and further research is necessary to ascertain whether those differences that were found were spurious in nature or were representative of subtle relationships among the variables.

The apparent ease with which observers changed their classification of slides is supportive of the applicability of a signal detection model of observation. Previously, the observational literature has argued that observers cannot readily be influenced to change their rating behaviors. This implication has been represented in the literature by the research on the resistance of behavioral data to bias effects (Kent et al., 1974; Siegel et al., 1976; Skindrud, 1972), as well as the more implicit assumption that training must take place over a long period of time during which observers practice often and gradually come to agree with each other or a standard. Contrary to this opinion, the present research suggests that brief instructions, similar to those used in classic signal detection and vigilance research, can produce changes in the rating behavior of subjects.

The effect of instructions coupled with the apparent differences among observers, as represented by the different biases of the subjects, lends further support to the use of psychophysical research as a model for the collection of observational data in the classroom. In typical observation research, attempts are usually made to expose observers to similar training experiences. However, if these individuals approach the standard task differently, then exposure to homogeneous experiences might be less effective than has previously been assumed (Johnson & Bolstad, 1973; Romanczyk et al., 1973; Wildman et al., 1975).

The most powerful support for the applicability of a signal detection model to the collection of observational data comes from the topography of the changes in recording, as well as the relationship between changes in the number of "True" responses and accuracy, as a function of the instructions. As signal detection theory would predict, subjects recorded more hits, but also more false alarms, under lenient instructions than under standard instructions. Conversely, subjects recorded fewer hits and fewer false alarms with strict instructions than they did with standard instructions.

Using a signal detection model, changes in accuracy would not be expected to accompany changes in the number of "True" responses; since as the number of correct "True" responses increases, so does the number of incorrect "True" responses, and, vice versa, as the number of incorrect "True" responses decreases, so does the number of correct "True" responses. Although the latter pattern occurred for subjects receiving strict instructions, accuracy scores declined significantly for subjects who received lenient instructions. These results may have been due to the way subjects were distributed, relative to the theoretical mean of 80 "True" responses. As discussed earlier, the preponderance of liberal subjects, based on a mean of 80 "True" responses, makes certain interpretations difficult. Since many of these subjects began with liberal rating tendencies, any shift to more lenient rating criteria would necessitate decreases in accuracy. However, the reverse is not necessarily true; strict instructions decreased the number of correct "True" responses concomitantly with decreasing the number of false alarms.

The results of the present research do not definitively establish signal detection theory as a model for the collection of observational data; however, they certainly suggest that signal detection theory is a viable model for the collection of these data. Such findings as the differences among individuals, the topography and relations between shifts in the number of "True" responses and accuracy, as well as the failure to find increases in the number of exact agreements are consistent with a signal detection analysis. Although each of the findings may be accounted for outside of the realm of signal detection theory, all of the results are consistent with a signal detection analysis. The goal of the present study was to assess the feasibility of tying together various aspects of observational research into a coherent theory, and the present research is supportive of using signal detection theory for this purpose. Additional support for the application of signal detection theory was reflected in the similarity of findings across slide conditions.

Clearly, this study was only the first of a series necessary to elaborate a theory of observational research and training. Important additions to the present research would include the replication of this study with a different population and different slides. Groups based on the biases of subjects should be formed relative to a theoretical mean in order to be able to make inferences that were difficult to draw from the present data. For brief instructions to be useful in training, the effects of various intensities of instructions should be investigated. More useful to observer training would be the exploration of the feasibility of employing instructions aimed at modifying an observer's classification of responses at both extremes of the slide distribution.

Finally, these principles would need to be applied to an array of behaviors, videotaped interactions, and finally to live situations. If signal detection and vigilance research are adequate models for systematic observation, procedures for maintaining consistent rating behavior, as well as training procedures may be improved. For example, vigilance studies have found that deterioration of rating over time can be alleviated by giving observers brief rests. Also interesting would be a comparison of records of observers trained in a conventional manner with the records of observers trained using brief individualized instructions.

BIBLIOGRAPHY

Arrington, R. E. Interrelations in the behavior of young children. Child Development Monographs, No. 8. Bureau of Publications, Teachers College, New York, 1932.

DeMaster, B., Reid, J., & Twentyman, C. The effects of different amounts of feedback on observer's reliability. Behavior Therapy, 1977, 8, 317-329.

Harris, R. J. Primer of multivariate statistics. New York: Academic Press, 1975.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts and practice. Champaign, Ill.: Research Press, 1973.

Kass, R. E., & O'Leary, K. D. The effects of observer bias in field-experimental settings. Paper presented at a symposium "Behaviour Analysis on Education", University of Kansas, Lawrence, Kansas, April 9, 1970.

Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. Expectation biases in observational evaluation of therapeutic change. Journal of Consulting and Clinical Psychology, 1974, 42, 774-780.

McNicol, D. A primer of signal detection theory. London: George Allen & Unwin, 1972.

O'Leary, K. D., & Kent, R. Behavior modification for social action: Research tactics and problems. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts and practice. Champaign, Ill.: Research Press, 1973.

O'Leary, K. D., Kent, R. N., & Kanowitz, J. Shaping data collection congruent with experimental hypotheses. Journal of Applied Behavior Analysis, 1975, 8, 43-51.

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

Siegel, L. J., Dragovich, S. L., & Marholin, D. The effects of biasing information on behavioral observations and rating scales. Journal of Abnormal Child Psychology, 1976, 4, 221-233.

Skindrud, K. An evaluation of observer bias in experimental-field studies of social interaction. Unpublished doctoral dissertation, University of Oregon, 1972.

Thomas, D. S., Loomis, A. M., & Arrington, R. E. Observational studies of social behavior (Vol. I). Yale University: Institute of Human Relations, 1933.

Wildman, B. G., & Erickson, M. T. Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, in press.

Wildman, B. G., Erickson, M. T., & Kent, R. N. The effect of two training procedures on observer agreement and variability of behavior ratings. Child Development, 1975, 46, 520-524.

Yelton, A. R., Wildman, B. G., & Erickson, M. T. A probability-based formula for calculating interobserver agreement. Journal of Applied Behavior Analysis, 1977, 10, 127-131.

Appendix A

Picture of Girl with 45° Head Orientation with White Background and Classroom Background

Pose 1

White Background

Pose 2

Classroom Background

Appendix B

INSTRUCTOR ______________ DEPT. ______________ COURSE ______________ ______________

STANDARD ANSWER SHEET

1 A B C D E	2 A B C D E	3 A B C D E	4 A B C D E	5 A B C D E	6 A B C D E	7 A B C D E	8 A B C D E
9 A B C D E	10 A B C D E	11 A B C D E	12 A B C D E	13 A B C D E	14 A B C D E	15 A B C D E	16 A B C D E
17 A B C D E	18 A B C D E	19 A B C D E	20 A B C D E	21 A B C D E	22 A B C D E	23 A B C D E	24 A B C D E
25 A B C D E	26 A B C D E	27 A B C D E	28 A B C D E	29 A B C D E	30 A B C D E	31 A B C D E	32 A B C D E
33 A B C D E	34 A B C D E	35 A B C D E	36 A B C D E	37 A B C D E	38 A B C D E	39 A B C D E	40 A B C D E
41 A B C D E	42 A B C D E	43 A B C D E	44 A B C D E	45 A B C D E	46 A B C D E	47 A B C D E	48 A B C D E
49 A B C D E	50 A B C D E	51 A B C D E	52 A B C D E	53 A B C D E	54 A B C D E	55 A B C D E	56 A B C D E
57 A B C D E	58 A B C D E	59 A B C D E	60 A B C D E	61 A B C D E	62 A B C D E	63 A B C D E	64 A B C D E
65 A B C D E	66 A B C D E	67 A B C D E	68 A B C D E	69 A B C D E	70 A B C D E	71 A B C D E	72 A B C D E
73 A B C D E	74 A B C D E	75 A B C D E	76 A B C D E	77 A B C D E	78 A B C D E	79 A B C D E	80 A B C D E
81 A B C D E	82 A B C D E	83 A B C D E	84 A B C D E	85 A B C D E	86 A B C D E	87 A B C D E	88 A B C D E
89 A B C D E	90 A B C D E	91 A B C D E	92 A B C D E	93 A B C D E	94 A B C D E	95 A B C D E	96 A B C D E
97 A B C D E	98 A B C D E	99 A B C D E	100 A B C D E	101 A B C D E	102 A B C D E	103 A B C D E	104 A B C D E
105 A B C D E	106 A B C D E	107 A B C D E	108 A B C D E	109 A B C D E	110 A B C D E	111 A B C D E	112 A B C D E
113 A B C D E	114 A B C D E	115 A B C D E	116 A B C D E	117 A B C D E	118 A B C D E	119 A B C D E	120 A B C D E
121 A B C D E	122 A B C D E	123 A B C D E	124 A B C D E	125 A B C D E	126 A B C D E	127 A B C D E	128 A B C D E
129 A B C D E	130 A B C D E	131 A B C D E	132 A B C D E	133 A B C D E	134 A B C D E	135 A B C D E	136 A B C D E
137 A B C D E	138 A B C D E	139 A B C D E	140 A B C D E	141 A B C D E	142 A B C D E	143 A B C D E	144 A B C D E
145 A B C D E	146 A B C D E	147 A B C D E	148 A B C D E	149 A B C D E	150 A B C D E	151 A B C D E	152 A B C D E
153 A B C D E	154 A B C D E	155 A B C D E	156 A B C D E	157 A B C D E	158 A B C D E	159 A B C D E	160 A B C D E

PRINT YOUR NAME IN THE BLOCKS PROVIDED. BLACKEN THE CORRESPONDING LETTER IN EACH GRID.

LAST NAME	FIRST	MI
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

USE A #2 PENCIL ONLY

1	2	3	4	DATE (MO, DAY, YR)	IDENTIFICATION NUMBER
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9

Appendix C

Table 1C

Multivariate Analysis of Covariance

for Lenient Instructions

Source of variance	df	F
Slide condition (C)	2	1.18
Subject bias (B)	2	1.42
C x B	4	2.60*
S (C x B) error	61	

* $p < .05$

Table 2C

Analysis of Covariance on Number of "True" Responses for Lenient Instructions

Source of variance	df	SS	MS	F
Slide condition (C)	2	461.39	230.70	0.82
Subject bias (B)	2	467.06	233.53	0.83
C x B	4	3025.63	756.41	2.68*
S (C x B) error	62	17484.76	282.01	

*$p < .05$

Table 3C

Analysis of Covariance for Number of

"Accurate" Responses for Lenient Instructions

Source of variance	df	SS	MS	F
Slide condition (C)	2	494.50	247.70	0.53
Subject bias (B)	2	153.61	76.81	0.79
C x B	4	1441.27	360.32	0.35
S (C x B) error	62	19752.29	318.59	

Table 4C

Multivariate Analysis of Covariance

for Strict Instructions

Source of variance	df	F
Slide condition (C)	2	1.20
Subject bias (B)	2	2.75+
C x B	4	0.65
S (C x B) error	61	

+$p < .10$

Table 5C

Analysis of Covariance for Number of "True" Responses for Strict Instructions

Source of variance	df	SS	MS	F
Slide condition (C)	2	472.90	236.45	0.81
Subject bias (B)	2	633.05	316.53	1.09
C x B	4	258.84	64.71	0.22
S (C x B) error	62	18048.32	291.10	

Table 6C

Analysis of Covariance for Number of

"Accurate" Responses for Strict Instructions

Source of variance	df	SS	MS	F
Slide condition (C)	2	202.70	101.35	0.53
Subject bias (B)	2	756.50	378.25	1.98
C x B	4	280.31	70.08	0.37
S (C x B) error	62	11850.95	191.14	

Table 7C

Multivariate Analysis of Variance for

Lenient versus Strict Instructions

Source of variance	df	F
Slide condition (C)	2	0.88
Instruction set (I)	1	75.55***
Subject bias (B)	2	39.96***
C x I	2	3.11*
C x B	4	1.75
I x B	2	0.59
C x I x B	4	2.14+
S (C x I x B) error	126	

+ $p < .10$

* $p < .05$

*** $p < .001$

Table 8C

Analysis of Variance for Lenient versus

Strict Instructions for Number of "True" Responses

Source of variance	df	SS	MS	F
Slide condition (C)	2	585.51	292.76	0.86
Instruction Set (I)	1	22927.01	22927.01	67.48***
Subject bias (B)	2	26901.51	13450.76	39.59***
C x I	2	1197.18	598.59	1.76
C x B	4	2355.65	588.91	1.73
I x B	2	362.26	181.13	0.53
C x I x B	4	1494.74	373.69	1.10
S (C x I x B) error	126	42810.63	339.77	

***$p < .001$

Table 9C

Analysis of Variance for Lenient versus Strict Instructions for Number of "Accurate" Responses

Source of variance	df	SS	MS	F
Slide condition (C)	2	42.76	21.38	0.06
Instruction set (I)	1	3916.67	3916.67	10.74***
Subject bias (B)	2	4609.43	2304.72	6.32**
C x I	2	1171.26	585.63	1.61
C x B	4	854.15	213.54	0.59
I x B	2	28.01	14.01	0.04
C x I x B	4	1958.74	489.69	1.34
S (C x I x B) error	126	45939.13	364.60	

** p .01

*** p .001

Table 10C

Multivariate Analysis of Variance
for Conservative Subjects Receiving Standard and
Lenient Instructions

Source of variance	df	F
Between		
Slide condition (C)	2	1.14
S (C) error	21	
Within		
Instruction set (I)	1	20.93***
C x I	2	2.50
I x S (C) error	21	

*** $p < .001$

Table 11C

Multivariate Analysis of Variance

for Average Subjects Receiving Standard

and Lenient Instructions

Source of variance	df	F
Between		
Slide condition (C)	2	2.13
S (C) error	21	
Within		
Instruction set (I)	1	16.40***
C x I	2	1.94
I x S (C)	21	

*** p .001

Table 12C

Multivariate Analysis of Variance for Liberal Subjects Receiving Standard and Lenient Instructions

Source of variance	df	F
Between		
Slide condition (C)	2	4.31*
S (C) error	21	
Within		
Instruction set (I)	1	45.89***
C x I	2	1.37
I x S (C) error	21	

* $p < .05$

*** $p < .001$

Table 13C

Analysis of Variance for Number of "True" Responses for Conservative Subjects Receiving Standard and Lenient Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	1169.29	584.65	0.94
S (C) error	21	13057.38	621.78	
Within				
Instruction set (I)	1	5002.08	5002.08	20.92***
C x I	2	1191.29	592.65	2.49
I x S (C) error	21	5021.63	239.13	

***$p < .001$

Table 14C

Analysis of Variance for Number of "True" Responses for Average Subjects Receiving Standard and Lenient Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	585.88	292.94	1.95
S (C) error	21	3161.44	150.54	
Within				
Instruction set (I)	1	1621.69	1621.69	13.97**
C x I	2	400.88	200.44	1.73
I x S (C) error	21	2436.94	116.04	

** $p < .01$

Table 15C

Analysis of Variance for Number of "True" Responses for Liberal Subjects Receiving Standard and Lenient Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	623.29	311.65	3.78*
S (C) error	21	1733.19	82.53	
Within				
Instruction set (I)	1	1989.19	1989.19	32.42***
C x I	2	166.63	83.32	1.36
I x S (C) error	21	1288.69	61.37	

* $p < .05$

*** $p < .001$

Table 16C

Analysis of Variance for Number of "Accurate" Responses for Conservative Subjects Receiving Standard and Lenient Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	69.88	34.94	0.02
S (C) error	21	30860.13	1469.53	
Within				
Instruction set (I)	1	1610.08	1610.08	5.74*
C x I	2	544.54	272.27	0.97
I x S (C) error	21	5890.38	280.49	

* $p < .05$

Table 17C

Analysis of Variance for Number of "Accurate" Responses for Average Subjects Receiving Standard and Lenient Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	559.54	279.77	1.83
S (C) error	21	3213.94	153.04	
Within				
Instruction set (I)	1	1271.02	1271.02	16.24***
C x I	2	147.04	73.52	0.94
I x S (C) error	21	1643.44	78.26	

*** $p < .001$

Table 18C

Analysis of Variance for Number of "Accurate" Responses for Liberal Subjects Receiving Standard and Lenient Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	1463.04	731.52	2.26
S (C) error	21	6785.94	323.14	
Within				
Instruction set (I)	1	910.02	910.02	6.99*
C x I	2	282.04	141.02	1.08
I x S (C) error	21	2734.44	130.21	

* $p < .05$

Table 19C

Multivariate Analysis of Variance for Conservative Subjects Receiving Standard and Strict Instructions

Source of variance	df	F
Between		
Slide condition (C)	2	4.67*
S (C) error	21	
Within		
Instruction set (I)	1	21.62***
C x I	2	0.86
I x S (C) error	21	

* $p < .05$

*** $p < .001$

Table 20C

Multivariate Analysis of Variance for Average Subjects Receiving Standard and Strict Instructions

Source of variance	df	F
Between		
Slide condition (C)	2	6.09**
S (C) error	21	
Within		
Instruction set (I)	1	37.76***
C x I	2	1.45
I x S (C) error	21	

** $p < .01$

*** $p < .001$

Table 21C

Multivariate Analysis of Variance for Liberal Subjects Receiving Standard and Strict Instructions

Source of variance	df	F
Between		
Slide condition (C)	2	2.34
S (C) error	21	
Within		
Instruction set (I)	1	9.35**
C x I	2	1.40
I x S (C) error	21	

** $p < .01$

Table 22C

Analysis of Variance for Number of "True" Responses for Conservative Subjects Receiving Standard and Strict Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	598.04	299.02	1.06
S (C) error	21	5932.88	282.52	
Within				
Instruction set (I)	1	1026.75	1026.75	5.09*
C x I	2	183.38	91.69	0.45
I x S (C) error	21	4232.88	201.57	

* $p < .05$

Table 23C

Analysis of Variance for Number of "True" Responses for Average Subjects Receiving Standard and Strict Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	1441.17	720.58	3.31+
S (C) error	21	1365.81	65.04	
Within				
Instruction set (I)	1	2422.52	2422.52	19.48***
C x I	2	95.17	47.59	0.38
I x S (C) error	21	2611.81	124.37	

+ $p < .10$

*** $p < .001$

Table 24C

Analysis of Variance for Number of "True" Responses for Liberal Subjects Receiving Standard and Strict Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	1210.79	605.40	2.34
S (C) error	21	5440.19	259.06	
Within				
Instruction set (I)	1	1017.52	1017.52	9.09**
C x I	2	122.79	61.40	0.55
I x S (C) error	21	2350.19	111.91	

** $p < .01$

Table 25C

Analysis of Variance for Number of "Accurate" Responses for Conservative Subjects Receiving Standard and Strict Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	5175.38	2587.69	3.82*
S (C) error	21	14218.88	677.09	
Within				
Instruction set (I)	1	2.08	2.08	0.01
C x I	2	428.04	214.02	0.79
I x S (C) error	21	5716.88	272.23	

* $p < .05$

Table 26C

Analysis of Variance for Number of "Accurate" Responses for Average Subjects Receiving Standard and Strict Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	122.00	61.00	0.94
S (C) error	21	1365.81	65.04	
Within				
Instruction set (I)	1	4.69	4.69	0.08
C x I	2	129.50	64.75	1.10
I x S (C) error	21	1231.31	58.63	

Table 27C

Analysis of Variance for Number of "Accurate" Responses for Liberal Subjects Receiving Standard and Strict Instructions

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	457.54	228.77	0.78
S (C) error	21	6143.44	292.54	
Within				
Instruction set (I)	1	1.69	1.69	0.01
C x I	2	294.88	147.44	1.18
I x S (C) error	21	2623.94	124.95	

Table 28C

Analysis of Variance on Arcsin Transformed Agreement

Percentages Using

$$\frac{\text{Agreements on Occurrences}}{\text{Agreements on Occurrences + Disagreements}}$$

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	0.00	0.00	0.00
S (C) error	21	6.67	0.32	
Within				
Instruction set (I)	1	0.23	0.23	1.51
C x I	2	0.17	0.08	0.55
I x S (C) error	21	3.18	0.15	

Table 29C

Analysis of Variance on Arcsin Transformed Agreement Percentages Using the Probability-based Formula

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	1.92	0.99	0.89
S (C) error	21	22.55	1.07	
Within				
Instruction set (I)	1	0.25	0.25	2.72
C x I	2	0.39	0.20	2.10
I x S (C) error	21	1.96	0.09	

Table 30C

Analysis of Variance on Arcsin Transformed Agreement Percentages Using

$$\frac{\text{Smaller Number of Occurrences}}{\text{Larger Number of Occurrences}}$$

Source of variance	df	SS	MS	F
Between				
Slide condition (C)	2	0.15	0.08	0.20
S (C) error	21	8.19	0.39	
Within				
Instruction set (I)	1	0.73	0.73	7.34*
C x I	2	0.11	0.06	0.57
I x S (C) error	21	2.09	0.10	

* $p < .05$

Table 31C

Mean Arcsin Transformed and Untransformed Agreement

Percentages, by Slide Condition, Using

$\frac{\text{Agreements on occurrences}}{\text{Agreements on occurrences + disagreements}}$, Probability-based Formula, and

$\frac{\text{Smaller number of occurrences}}{\text{Larger number of occurrences}}$

	agreements / agreements +disagreements	transformed agreements / agreements +disagreements	Probability-based formula	transformed probability-based formula	Smaller / Larger	transformed smaller / larger
Lab slides						
Standard instructions	0.62	1.83	0.12	0.36	0.67	1.86
Experimental instructions	0.62	1.54	0.23	0.74	0.78	2.24
Classroom without sound slides						
Standard instructions	0.59	1.76	0.00	0.01	0.61	1.82
Experimental instructions	0.61	1.63	0.03	0.12	0.69	2.01
Classroom with sound slides						
Standard instructions	0.56	1.70	0.10	0.28	0.64	1.87
Experimental instructions	0.63	1.70	0.07	0.22	0.70	2.03