THE UNIVERSITY OF MANITOBA

SOCIAL CONTAMINANTS OF OBSERVERS USED IN THE STUDY OF NONVERBAL BEHAVIOUR

by

Robin Douglas Peace Montgomery

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

WINNIPEG, MANITOBA

October 1971

ACKNOWLEDGEMENTS

I wish to express my deep appreciation to Dr. Marion Aftanas, Dr. Hugh McGinley, and Dr. Daniel Perlman for their interest in this investigation and for their many helpful suggestions. Especially I wish to thank Dr. John Adair for the fellowship of inquiry in which these studies developed, for the innumerable insights afforded by his guidance, and for the patience with which he directed the preparation of this dissertation from its disorderly birth to its present form. I would also like to thank my wife and family for the years they have been willing to be without me in the interests of psychological research.

ABSTRACT

This investigation explored the contaminating influences which appear to affect observers used to record the nonverbal responses of subjects in psychological experiments. The problems which may arise are serious since observers are widely used for this purpose, and are sometimes the only means of obtaining the data, and because the danger of observers being susceptible to such influences has been largely ignored.

An historical survey of the use of the observer technique reveals that investigators have been generally more concerned with the technical aspects of the problem than with the social psychology of the observer. A few studies have shown that observers' expectancies may differentially influence their scoring. Apart from rare theoretical speculations it has been assumed that the influence of paired observers upon each other is not a matter for concern. It has also been assumed that high observer agreement, as measured by reliability coefficients, is an indication of objective scoring. Consideration has not been given to the possibility that it may be due to inter-observer influence. Two aspects of observer contamination were therefore studied in this investigation; the effects of observers' expectancies on their scoring and of inter-observer influences.

Observers participated in a practice session followed by a testing session a week later. They scored the smiles and nods of several subjects presented on identical videotape recordings to all observers. Expectancies were manipulated by telling different groups of observers

that the subjects viewed in the test session liked or disliked the person interviewing them, or were neutrally disposed towards him. Results indicated that the observers' expectancies influenced their scoring when they scored alone. The observers were examined by three different procedures to determine inter-observer influences. Some observers scored alone in both practice and test sessions. Others were paired in the first session and scored alone in the second. Observers in a third group were paired in both sessions. The influence of paired observers upon each other was assessed by comparing the level of scoring agreement between partners with that of the same sets of observers randomly matched and with that of observers who had scored alone. Inter-observer influence was demonstrated when observers scored smiles, paired in both practice and test sessions, and when they scored nods, paired in the practice session but alone in the test session, the strength of these effects being related to the conditions of observer expectancy. In a second study an attempt was made to manipulate inter-observer scoring consensus by pairing the observers with confederates who scored at a consistently high or low level. The effect was demonstrated for smiles when the observers were paired with the confederates in both practice and test sessions. It was demonstrated for nods when observers were paired with confederates in both sessions and also when they were paired only in the practice session, the effect being stronger in the latter condition. Thus, the results of the second study not only strengthened the conclusions that may be based on those of the first, but indicated that inter-observer consensus can be demonstrated by experimental manipulation.

It was concluded that the observer technique is not free from contaminating influences, as had generally been supposed. Specifically,

contamination due to observer expectancies and to observer consensus has been demonstrated. These findings raise serious questions regarding studies in which observers are employed. Not only must precautions be taken in such studies to eliminate these subtle influences, but much intensive investigation is required.

TABLE OF CONTENTS

LIST OF TABLES

Table | Page

CHAPTER I

INTRODUCTION

The purpose of this study was to examine several potential sources of contamination in observational techniques used in the study of human behaviour. Because human observers are involved it is difficult to prevent the operation of inter-personal influences and, consequently, the contaminating effects of such influences upon their observations. While contaminating effects have been documented as they relate to the interactions between experimenters and subjects (e.g., Orne and Scheibe, 1964; Rosenthal, 1966), little attention has been directed to similar contamination in observers. However, since the observer technique is widely and frequently used, the possibility of widespread research contamination requires investigation. An understanding of the source of contaminants will enable investigators to control them or, at least, to allow for their effects.

Observation, as it is used in these psychological investigations, is the process by which a person attempts to record the responses of one or more subjects in accordance with the instructions provided by the experimenter. The task of observation is a composite one in which extraneous variables may affect the observer's perception, the cognitive and affective processes to which the perceived response is subjected, and the recording of his final judgment.

While the wide and frequent use of observational techniques justifies concern regarding the effects of observers' susceptibility to contamination,

the fact that these techniques are almost the exclusive source of information in many areas renders concern even more imperative. This is especially true in the investigation of nonverbal communication. Since the subject is usually unaware of performing the behaviour that is being scored and is liable to be selective in his recall, and since the subtle nature of the responses often precludes data collection purely by mechanical scoring apparatus, these main alternative sources of data, namely, the subject himself or the mechanical scoring of responses, can not be utilized in many experimental situations. Thus, while an investigator may have some suspicions that various sources of social contamination are associated with the use of observational techniques, the absence of a viable alternative may lead him to repress his doubts about the validity of these techniques.

In order to provide guidance for researchers who find themselves in this predicament there is a need for investigations which explore the technique and which clarify the nature and effects of these social contaminants. While a given study may examine only one social contaminant, ultimately the concern is with all influences upon observers which lead to results that are due to variables other than the responses as specified in the scoring instructions. Furthermore, investigations of this nature must also take into account the fact that "error" in these observational techniques can not be examined in relation to any veridical standard. The only possible standards of comparison are the scores of other observers.

This study, therefore, examines several sources of systematic social contamination which seem to be present in studies employing the observational technique. One major concern is with the contamination

due to attitudes and hypotheses aroused in the observers by their pre-observation expectations, an effect somewhat similar to the influences of experimenter expectancy upon subjects (Rosenthal, 1969). A second major concern is the contaminating effects of the inter-personal nature of the scoring procedure where two observers score the responses together This situation is likely to provide opportunities for nonverbal communication between the observers, resulting in the scoring being influenced by interobserver consensus.

CHAPTER II

HISTORICAL REVIEW OF OBSERVER CONTAMINANTS

Throughout the many years that observational techniques have been used in the study of human behaviour authors have, from time to time, drawn attention to various problems of methodology and interpretation. However, there has been little appreciation of the implications of these problems, as they affect the use of observers in research, and they have been almost ignored as a subject of investigation.

The lack of research reports dealing with the observer as a tool to record data is referred to by Heyns and Lippitt (1954) in their review of observational techniques. They note that

> There is a relative paucity of information concerning factors which affect the accuracy and reliability of observer data. Strictly methodological studies involving the use of human observers are rare (p. 370).

Smith (1966), in a volume dealing with sensitivity training, expressed similar dissatisfaction with the state of empirical investigation of observation. "Psychologists have shown a curious lack of interest in this determinant (p. 19)."

It is true that observation has been studied empirically. This applies particularly to the process of observation itself, the technique of participant observation, and also the halo effect. While important contributions have been made in these areas such studies do not specifically investigate the observer in the experimental situation.

Investigations of the process of observation by Sherif and Asch

have shown that inter-personal influences may contaminate observers. Sherif (1936) found that subjects observing a stationary point of light, in an otherwise completely darkened room, reported that the light moved. After some practice these subjects were able to give fairly consistent estimates of the distance of the apparent movements. However, when subjects who had formed their own norms of responding under conditions of individual testing were exposed to the influence of group members who expressed their judgments orally, they tended to approach a norm peculiar to the group. A series of somewhat similar studies was conducted by Asch (1951). Instead of using a stationary light in a darkened room as stimulus, subjects were instructed to select from several lines of different lengths the one that corresponded to a line of standard length. In the crucial trials a group of confederates unanimously chose an obviously wrong line and made their choice known to the subject before he had expressed his own decision. Although many subjects did not respond to the influence of the confederates, there was a strong tendency for the responses of some observers to be distorted in the direction of the judgments expressed by the majority. Sherif and Asch have thus demonstrated that observers' responses are contaminated in such situations by a tendency to deviate toward what they believe to be group accord.

The possibilities of contamination are so obvious when participant observers are used that here also some doubts have arisen as to the validity of scores obtained by this method. For example, Vidich (1954), in discussing the social position of the participant observer, points out that the respondent's image of the observer is a basis for his response, that the participant observer must react to his respondents in order to maintain his role, and that the observer's category boundaries change as

a result of his increasing experience in the observational situation. Schwartz and Schwartz (1954), commenting on the participant observer mainly as he is used in anthropological situations, raise other concerns. They discuss such effects as the contamination due to the observer's retrospective activities between the moment of input and the recording of the response, the danger of the observed "producing" data for the participant observer, and the loss of objectivity if the observer becomes emotionally involved in the situation. It is evident that some of the issues raised in these discussions have their source in the participant nature of the observation while others are inherent in any process of observation. It is also evident that many of the observational problems which should be cause for concern in the more structured research situations of the laboratory are present as well in participant observation. While some of the issues investigated in the present study are common to both participant observation and to observer techniques as used in the laboratory, the concern of this study is limited to observer contamination as it occurs in the laboratory.

One of the more serious problems associated with observation has been particularly noted in studies of the halo effect. Thorndike (1920) drew attention to the surprisingly high correlations between the traits of a person rated by any one observer. The halo effect has since been demonstrated by experimental manipulation (e.g., Johnson and Vidulich, 1956). Such studies have shown that a rater is influenced in scoring a particular trait by his general impression of the ratee rather than by attributes exclusively related to the particular trait. It would seem to be a plausible conclusion that the scores of observers would be contaminated, like the scores of raters, by the observers' general impression

of the subjects they are scoring.

Observer Expectancy Effects

There is some evidence to indicate that observers may develop predispositions or biases which systematically affect their scoring, in a manner analogous to the halo effect. For instance, an observer scoring a particular category of response may be influenced by his general impression of the subject he is scoring. Even differences between the subjects created by the various experimental treatments may produce different expectations in the observers and thus lead to biased scoring. It has in fact been demonstrated in experimental investigations that experimenters' results can be biased by telling them that high or low scores are anticipated (Rosenthal, 1966). Similarly, it may be expected that the observer's expectations will influence his output.

One of the most explicit statements of this problem is that provided by Rosenthal (1966) in his volume dealing with experimenter effects. After discussing the probability that even experienced observers may differ in their perception of behaviour he continues, "A somewhat different but perhaps more serious problem, however, is that in which observer effects interact with experimental conditions (p. 8)." As evidence he refers to the study by Rosenthal and Halas (1962). In this investigation two experimenters attempted to condition planaria to respond, by head turning and body contraction, to a light which had been paired with an electric shock. One of the experimenters obtained no significant effects for either turns or contractions. The second experimenter obtained a significant increase in head turning in the experimental group, the animals showing increases over successive blocks of trials. The same experimenter also obtained a significant increase in body contractions in

the experimental group but an even greater increase in contractions in the control group. This is particularly remarkable as previous research had shown the turning and contracting responses to be so highly correlated that they have been commonly added together to form a total response score. Although Rosenthal suggests possible alternative explanations he rejects them because of their improbability and claims that

> It can be concluded that there are individual differences in the extent to which behaviour modifications in planaria are observed and that the particular differences found are affected by the specific type of behaviour being observed (Rosenthal, 1966, p. 9).

A somewhat similar effect was produced by Cordaro and Ison (1963) who manipulated the expectations of observers by casual statements just prior to the experimental session. Nearly five times as many head turns, and twenty times as many body contractions, were reported by observers who expected high levels of responding, in comparison with those who expected low levels. Cordaro and Ison concluded, "It is quite clear that O's expectations have a large effect on the incidence of reporting responses (p. 789)."

Bias resembling the effects of experimental conditions upon observers discussed by Rosenthal has been found in a different area by Dudycha and Naylor (1966). These investigators examined the effects of experimental conditions on raters' judgment policies. Specifically, they were concerned with whether the judgment and scoring of the raters depended partly on their conception of the experimental conditions, and whether a change in these conditions would predispose the raters to score differently. The experimental task was to rate overall job desirability by examining job profiles that indicated the degree to which each of the several traits was present in each job. Three experimental conditions

were employed; actual profiles, artifactual profiles in which the job traits had a zero relationship with each other, and artifactual profiles in which the job traits had a relatively high inter-relationship. Each profile was displayed to the rater for approximately 12-15 seconds. An analysis of variance revealed a significant Conditions effect, indicating that "the policy levels for the three conditions were significantly different (p. 595)." The authors also concluded that a significant interaction between Job Traits and Conditions indicated that "dissimilar policies emerged from the three conditions, that is, the job traits received different relative emphasis within the three conditions (p. 595)." While these judges did not observe human subjects the abstract nature of the stimuli suggests that the biasing effect is likely to generalize to all tasks requiring human judgment, including tasks of observation in investigations of nonverbal communication.

In another area, the diagnosis of hearing defects in neonates, Ling, Ling, and Doehring (1970) have investigated the influence of observers' expectations on their scoring. Two observers recorded several categories of responses to auditory stimuli of neonates up to six days old. One observer could hear the stimuli but the other was prevented from hearing them by a masking noise. It was found that masked observers scored significantly more false-positive responses, i.e., responses recorded when no sound stimulus had been given, than did unmasked observers. This study may have been fully adequate for its primary purpose, the investigation of certain diagnostic procedures. That is, for the study of the particular clinical situation it may be immaterial whether errors originate from the neonate responding when no stimulus is given or from the observers scoring the responses inaccurately. In so far as its findings relate

to the effects of observer expectations, however, the conclusions have an element of ambiguity. It seems unlikely that a neonate would never emit any of the several responses studied except when an experimental stimulus was presented. If such unintended responses did occur it would be difficult to determine whether it was the masked or unmasked observers, or either, who scored objectively. Nevertheless, while there may be doubt as to which set of observers was responsible for the deviation in scoring, it is clear that the procedure of using masked and unmasked observers did result in differences in scoring, presumably through the influence of these manipulations on the observers' expectations.

In addition to these empirical demonstrations of observer bias, the matter has been raised in theoretical discussion. Pawlicki (1970), in a critical review of behaviour-therapy research with children, points out that high inter-observer reliability correlations should not be misconstrued as a substitute for unbiased observers. He claims that, in 96 per cent of the 50 studies reviewed, precautions were not taken to ensure unbiased observers, and suggests the use of videotape recordings to eliminate the transmission of experimental treatment cues to the observers.

In conclusion, it appears that observers may be contaminated by the hypotheses or expectations they formulate regarding the subjects they observe, including their hypotheses and subsequent expectations related to the subjects' experimental conditions. Since such contamination could produce a confounding of the direct effects of the experimental treatments on the subjects with the indirect effects of these treatments on the observers it is evident that a detailed examination of the problem is necessary. Such an examination is one of the major concerns of this study.

Veridicality of Observers' Reports

It is apparent from the studies outlined above that one of the major problems in assessing the performance of observers is the lack of any veridical standard of comparison. If some objective method of recording, that did not depend on human observers, was available it would obviously be used. The fact that human observers are used implies that no better technique is available. Unfortunately, human observers are not only subject to the biasing influences already discussed but also reveal wide individual differences in their scoring. This was demonstrated by Allport (1924) in investigating ability to judge facial expressions in pictures. Scores ranged from 21 to 72 per cent accurate in relation to the investigator's assumed standard of accuracy.

Not only do observers often fail to record accurately the variations in external stimuli, but they may be unable to report accurately their own experiences, as is documented by Syz (1926). Stimulus words, estimated to arouse emotional responses of various intensities and social acceptability, were read to subjects and the latter's galvanic responses were compared with their verbal reports of affect. It was found that verbal reports corresponded with galvanic responses when mild affect was aroused but that there were marked discrepancies, between the two modalities of responding, for words that aroused intense and unacceptable emotions. While unquestionably there are differences between self-observation and observation of the external world, Syz has demonstrated that, given the input, contaminating influences can seriously distort the observers' output.

A few authors, who have been concerned with the contaminations of the observer in laboratory situations, report their endeavours to remedy the difficulties. Bernhardt, Millichamp, Charles, and McFarland (1937)

attempted to get more objective data by having observers score responses from live behavior and from filmed recordings of the identical behavior and then combining the two sets of scores. They also combined scores obtained through repeated observation of a film by one set of observers. It was claimed that these procedures generally expanded the information obtained without changing the overall picture of the behaviour.

In addition to the use of movie recordings attempts have been made to control for lack of veridicality and for systematic human error in observation by the development of mechanical scoring devices (e.g., Chapple, 1949; Haith, 1966). Most of these devices, however, did not eliminate the human observer entirely and some, by their nature, were restricted to scoring very molecular responses.

Other investigators have approached the problem by providing some statistical estimate of the quality of the observer's performance, in the hope of thus removing concern for the possible inadequacy of the observational data. For example, multiple observers presumably provide a check on the inadequacies of single observers. Reliability coefficients have been employed, as affording some statistical indication of the degree of agreement between the scores of different observers, and the acceptability of the scores has been decided largely on the basis of such agreement.

There can be no doubt that a correlation of the scores of paired observers provides a measure of the observers' reliability, in the statistical sense of the term, and that such reliability is a <u>necessary</u> condition of validity. It is not, however, a <u>sufficient</u> condition of validity. Such correlational estimates can indicate, by low correlation coefficients, that there is a lack of agreement between the observers in their scoring. A finding of this nature should lead the investigator to

conclude that the scoring performance of his observers is unacceptable. However, a high correlation, indicating high agreement between the observers, would not justify the conclusion that their scoring has reached acceptable standards. While reliability has been established in this case, such results do not indicate anything more than consistent agreement between observers. They are certainly not an indication of the validity of the scores, although it is easy for investigators to imply that they are (e.g., Exline, 1963; Rosenfeld, 1966; Mehrabian and Williams, 1969). In an extreme situation, for example, the observers might all have mistaken the instructions and scored the wrong response. It is conceivable that such scores would yield extremely high reliability coefficients, but they would obviously not be valid.

Observer Pairing Effects

A further possibility which can not be ignored is that high reliability coefficients may be due to agreement among observers based merely on consensus. This could be attained through an exchange of scoring information between observers by means of nonverbal cues. In fact Campbell (1958), in his review of sources of systematic error due to human links in communication systems, explicitly states that such consensus will occur "when a group of persons are exposed to the same input and are asked to transmit or code it, if they are in communication among themselves (p. 361)."

While the inadequacy of measures of reliability as indicators of validity is not open to question, the possibility of observers being contaminated by consensus is an issue which can only be decided by a detailed empirical investigation. The results of an earlier study (Montgomery and Adair, 1971) supported the notion that agreement between

paired observers may be, at least in part, due to such consensus. In that study it was found that while the Pearson product-moment correlation for the scores of paired observers yielded coefficients significantly above zero the corresponding coefficients for the same data when the observers were matched randomly were approximately zero or were negative. It is assumed that, if the scores of partners are in relatively high agreement but the scores of the same observers randomly matched within experimental conditions show significantly lower agreement, the former measure must have been spuriously inflated by consensus between partners. Otherwise, if the relatively high agreement between partners was due to the satisfactory scoring performance of the observers, the level of agreement would not be appreciably affected by the process of random matching. The concern of the present study includes a partial replication of the earlier investigation with actual manipulation of the observer's training and "testing" environment.

Classification of Observer Errors and Deviations

The foregoing review suggests that there is a number of sources of contamination. For the purposes of this study, and for clarity in an overview of the contaminating influences in the observer technique, it may be helpful to define some broad distinctions between categories of observer contamination. Rosenthal (1966), in discussing observer problems, has informally categorized observer errors into a dichotomy, namely, random errors on the one hand, and systematic errors arising from observer bias on the other. Further examination of the nature of observers' scoring errors suggests that categorization into the following trichotomy is more satisfactory.

Non-systematic scoring errors. Such errors, presumably having a dispersion comparable to that of normal discriminal judgments around the mean, are simply classified as observer errors. These errors may be due to the observers' inability to attain a greater precision of judgment, or to such external factors as sudden interfering noises from outside the observation room. Observer errors can usually be assumed to be largely self-cancelling, and no serious consequences result unless the magnitude of the errors is great. They are therefore assumed to be random in this investigation.

Observer expectancy effects. Systematic scoring deviations which result from the observers being influenced by the experimental treatments administered to the subjects may be classified as observer expectancy effects. The effect is dependent upon the observers forming hypotheses, true or false, regarding the experimental conditions of the subjects and being influenced differentially in their scoring by these hypotheses. For example, if an observer believes that one experimental group has received electric shocks and that another group has received no shocks, his resulting expectations may predispose him to score more frequent displays of aggression for the former group than for the latter, irrespective of the actual frequency of their aggressive responses. Similarly, if an observer concludes that a particular group of subjects liked an interviewer, and that another group disliked the interviewer, these conclusions may produce corresponding liking and disliking for the two groups of subjects on the part of the observer. Such observer expectancies could be reflected in experimental results by the data varying differentially over the experimental conditions. As stated earlier, observer expectancy effect is one of the concerns of this study.

Observer effects. Systematic inflations or reductions of scores, due to constant influences other than observer expectancies upon the observers, may be classified as observer effects. For instance, all scores may deviate in the same direction because the observers have a high need for the approval of the investigator and believe that his approval can be attained by recording a large number of responses. Again, an awareness that a companion is scoring more, or less, frequently than himself may cause an observer to raise or lower his own level of scoring to correspond with that of his partner. The principal distinction between observer errors and observer effects is that the former are random while the latter are systematic. The principal distinction between observer expectancy effects and observer effects is that the former is related to the outcome expectations of the observers while the latter are related to such factors as traits of the observer or his knowledge of how his companion is scoring. One aspect of observer effects, i.e., scoring consensus attained by paired observers through an exchange of nonverbal scoring cues, is investigated in this study.

Statement of the Problem

Observers are frequently used in experiments in social psychology and their use is often essential in studies of nonverbal communication. However, problems arise in the use of the observer technique because of the difficulties in assessing the objectivity of their scoring. Inter-rater reliability, as measured by correlating the scores of different observers, can not be assumed to be a measure of the validity of the scores. Furthermore, scoring consensus between observers, resulting from a nonverbal exchange of scoring cues, may render these reliability

measures inappropriate for assessing the performance of the observers. A further problem, of equal importance in relation to the observer technique, has emerged from the study of the experimenter expectancy effect. It has been found that observer expectancies can contaminate and invalidate results. The purpose of this study is, therefore, to examine and clarify these problems in inter-observer consensus and observer expectancy effects as they affect the observer technique.

As a basis of one aspect of the investigation it is speculated that the consensus variable can be measured by comparing the agreement of observers who participate as partners with the agreement obtained from the same scores of these observers randomly matched. The former measure of agreement is expected to be greater than the latter, the magnitude of the difference being an indication of the degree of consensus between the partners. A control group of observers scoring alone is also employed with a view to demonstrating that agreement between members of this group would be lower than that of the partners, inflated by consensus, and higher than that of the randomly matched partners whose agreement would be reduced by consensus contamination. A second control group, in which observers practice in pairs and score alone in test trials, is used to eliminate the possibility that agreement between partners is due to the observers reaching consensus on response definitions. It is assumed that if such consensus occurred in practice it would result in higher agreement between members of pairs even when they later score alone in test trials, since the observer's change in response definition would presumably be of an enduring nature.

Observer expectancy effect is investigated by manipulating the information supplied to the observers regarding the experimental conditions of

the target subjects. The observers' scoring is expected to vary in accordance with the expectancies thus aroused.

A further study is conducted for the purpose of demonstrating that the consensus effect can be manipulated experimentally. The major difference between this study and the first is that, here, instead of having a real observer as partner each observer is paired with a confederate posing as an observer and scoring consistently at a predetermined high or low level. It is hypothesized that the scores of the observers will be high or low in conformity with the scoring levels of the confederates with whom they are paired.

Specifically, therefore, it was hypothesized:

1. That the number of smiles and nods scored will be a positive function of the favourability of the ostensible attitude of the target subjects towards the interviewer.
2. That correlations for the scores of observers who were paired in both sessions will be significantly greater than zero.
3. That correlations for the scores of all randomly matched observers, and of observers who were paired in the first session and scored alone in the second, will not differ significantly from zero.
4. That correlations for the scores of observers who were paired in both sessions will be significantly greater than the correlations for the scores of the same observers randomly matched.
5. That correlations for observers scoring alone in both sessions, and for observers who were paired in the first session and scored alone in the second, will be intermediate in value between the correlations for paired and for randomly matched observers who scored with partners in both sessions.

6. That results for difference scores will follow a pattern similar to that predicted for the correlational results. That is, difference scores for observers paired in both sessions will be significantly smaller than the difference scores of the same observers randomly matched, and difference scores for observers scoring alone in both sessions, and for observers who were paired in the first session and scored alone in the second, will be intermediate between the difference scores for paired and for randomly matched observers who scored with partners in both sessions.

7. That, in the second study, scores of observers who were paired with a high-scoring confederate will be significantly greater than the scores of those paired with a low-scoring confederate.

CHAPTER III

METHOD

Study I

Observers

The subjects (Observers) were 144 male students from the Introductory Psychology courses at the University of Manitoba, who were fulfilling requirements for participation in laboratory experiments. They were allocated randomly to the various experimental conditions and took part in two experimental sessions, of one hour duration, at one week intervals.

Twenty-three observers were eliminated through their failure to show up for one of the two observational sessions. A further three were eliminated because they did not draw the line across the scoring paper when the signal was given to do so, thus failing to carry out experimental instructions. The eliminated observers were replaced by an equal number of observers from the same population.

Experimental Design

There were three conditions of observer expectancy and three conditions of observational pairing. The three conditions of Expectancy were differentiated by whether the observer was told that the target subject in the videotape he was viewing liked (L) or disliked (D) the person interviewing him, or was neutrally disposed (N) and neither liked nor disliked him. Forty-eight observers participated in each of these conditions of Expectancy.

Conditions of Pairing differed according to whether the observer scored alone or in the company of a partner in each of the two observational sessions. In the Alone-Alone (AA) Condition 48 observers, 16 in each of

the three conditions of Expectancy, participated alone. That is, no other observers were present for the training or test sessions. In the Pairs-Alone (PA) Condition 48 more observers, also allocated in equal numbers to the three conditions of Expectancy, scored as pairs in the first session and alone in the second. A further 48 observers, likewise allocated in equal numbers to the three conditions of Expectancy, scored in the Pairs-Pairs (PP) Condition in which the same pair of observers participated together in both sessions. Thus there were nine experimental conditions: (AA-L), (AA-N), (AA-D), (PA-L), (PA-N), (PA-D), (PP-L), (PP-N), (PP-D).

Preparation of the Stimulus Materials

Basically the stimulus materials consisted of ten-minute videotape recordings of each of five target subjects as they were being individually interviewed. Recordings were made, unknown to the target subjects, while they answered a number of general questions presented on a tape recorder by a confederate. The hidden camera was located approximately eight feet from the target subject at an angle of about 45^{o} from the direct front view. A picture of good resolution, in which the subject's movements were clearly visible, was obtained. All target subjects were seen from approximately the knees up.

While it was realized that the target subjects might have differed individually in appearance and behaviour, the concern in this investigation was with the comparison of observers who received differential experimental treatments. Since each observer saw the same target subjects in the same sequence, differences between target subjects remained constant for all observers.

The Observation Situation

All instructions were also presented on videotape, to ensure that each observer received a consistent instructional set. For these sound-videotape recordings a fourth-year honors student was directed to read the instructions in a friendly relaxed manner, approximating the reading of instructions in a typical experimental situation. By this method it was possible to ensure that the presentation was not altered to accommodate any particular observer.

In the first, or practice, session these videotapes provided all observers with task instructions. The content of the instructions varied, depending on the experimental conditions, and is described in the Procedure section which follows. In all, five sets of videotapes were required. Two were used for the first session, one for observers scoring alone and one for observers scoring in pairs. Each of these tapes also contained the interviews of the three target subjects to be used for practice. For the second or test session the other three tapes were prepared. Each of these presented the manipulation for one of the levels of Expectancy and each contained the interviews for the final two targets subjects.

The videotape recordings were presented to the observers on a Sony Videocorder TCV-2010, which included a television set with a nine-inch monitor. Observers sat approximately five feet nine inches from the monitor and four feet from each other. The observers' chairs and the television set remained at constant marked positions for all trials.

Responses were scored by pencil on adding-machine paper rolls inserted in simple dispensers. The table at which the observers scored was approximately seven feet long so that members of observer pairs

could be seated several feet apart, to limit opportunities for nonverbal communication. As a further precaution against such communication a partition, approximately 21 inches high, was placed at a point midway between the ends of the table. This partition was employed for all trials in which observers scored in pairs. For trials using single observers one chair, in alternating rotation, was removed from the table beforehand. Immediately below the television set there was a sign, in large letters, "X = SMILES O = NODS." This sign was intended to prevent the observers from confusing the scoring symbols during observational trials.

Procedure

First observational session. The experimenter escorted the observers from the waiting room to the observation room where he directed them to be seated in the observers' chairs. On those occasions when only one of the two observers arrived for the first session the experimenter made the necessary changes for single observer participation before conducting the observer to the observation room. The observers were then requested to attend to the television screen, on which all instructions were presented. When the videotape began to play the experimenter seated himself behind the observers' range of vision and the instructor appeared on the screen.

The instructor welcomed the observers and assured them that they would not be used as subjects but as observers. He then explained how the target subjects had been recorded and the nature of the observers' task. The full text of the instructions is presented in Appendix A. As a cover story the observers were told that the target subjects had worked with the interviewer in the construction of a simple crossword puzzle, "a task that drives most people either to enthusiastic co-operation or to

bitter disagreement." It was explained that the subjects were subsequently divided into those who liked or disliked the interviewer or were neutrally disposed, and that the effect of these different conditions on the frequency of the subjects' smiles and nods was being investigated, however they were not told at this time what to expect in their observations. The observers were then given directions in scoring procedure, being told to score an "X" on the paper tapes whenever the subject smiled and to score an "O" whenever he nodded, to pull the paper towards them as required, and to keep focussed on the subject and not look down at the paper while observing. A nod was defined both verbally and by the instructor illustrating the movements with his head. It was considered unnecessary to define a smile. Observers were told that some practice would be necessary to enable them to attain accuracy in scoring. It was explained that a bell would be rung halfway through each trial, and that, on hearing this signal, they should draw a line across the paper tape to permit an investigation of changes in responding over time. Their attention was drawn to the marks on the floor, designed to ensure that the position of the observers' chairs would not vary. Paired observers were told that the purpose of the partition was to prevent inter-observer influence during scoring, and the importance of independent scoring was emphasized. The observers were also requested not to talk while observing the subject. These references to inter-observer communication were, of course, omitted when observers scored alone. A brief summary of the instructions was presented and the observer was told, should the instructions have left him in any uncertainty, to use his best judgment. The instructor appealed to him to do his best to score accurately. Finally, the observer was asked to write his name, and the number of the subject, on the tape

and told to prepare to focus on the subject.

The first subject was seen on the screen for ten minutes and his responses recorded by the observers. Halfway through the trial the experimenter rang the bell. The main purpose of the bell was to break the monotony of the ten-minute trial and to help maintain the alertness of the observers. Since these measurements, the differences in the frequency of the responses over the two halves of the trials, had not proved of any utility in two previous studies, no predictions were made, although the data were examined according to each separate time interval.

Immediately after the first subject had been scored the instructor reappeared on the screen. He reminded the observers to draw a line across the paper tape, beyond the last score, and to tear it off beyond the line. The observers were then allowed to rest for 30 seconds. During this and all rest periods the screen remained blank. At the end of the 30 seconds the observers were asked to check that their chair legs were at the marks on the floor, a procedure which was employed at the beginning of every trial in both sessions. The procedure followed for scoring the responses of the first subject were then repeated for the second. Since fatigue presumably increased as the session progressed, observers were allowed to rest for one minute, rather than 30 seconds, between scoring the second and third subjects. The procedure already employed was then applied to the scoring of the third subject. After the third trial had been completed the instructor thanked the observers for their services and reminded them that they were due to return for the second session a week later. He also made the suggestion that, as a further reminder, one observer phone the other on the evening before the second session.

Instructions for observers scoring alone in the first session (AA Condition) did not differ from the above, apart from the omission of references to inter-observer communication, to the partition, and to phoning as a reminder prior to the second session. The partition was not used, under AA treatment, and one of the chairs was removed from the table before the observer arrived.

Observers scoring under PA Condition received the same instructions as those described above for PP Condition. At the end of the session the experimenter arranged for the members of each pair of observers under PA Condition to return at separate times for the following session, the second observer returning half an hour after the first.

Second observational session. The general procedure was similar to that employed in the first session. Each observer remained under AA, PA or PP Condition according to which of these treatments had already been administered to him. However different attitudes towards the target subjects were induced by instructions describing the target subjects as liking, disliking, or being neutrally disposed towards the interviewer.

Observers were again welcomed by the instructor and were reminded briefly of the nature of their task. After they had written their names and the number of the subject on the paper tapes they were told, as cover story, that it had been found that accuracy of scoring is increased if observers knew the target subject's emotional state, which in this study would be his disposition of liking, disliking, or neutrality toward the interviewer. Consequently, it was explained, the observers would be told each target subject's condition. They were also instructed to write the target subject's condition on the paper tape, beyond his number, and to

try to remember his condition throughout the trial. Observers with the Liked Expectancy were told, immediately before scoring each of the remaining two target subjects, that the subject liked the interviewer, and were instructed to write the word "Liked" on the paper tape. Apart from serving to identify the tape, it was hoped that writing the supposed condition of the target subject would help to impress the information on the memory of the observer. This procedure was repeated for both the Neutral and Disliked Conditions. The two test subjects were then scored, using the procedure employed in the first session with the modifications that have been described. All observers saw identical recordings of the target subjects, regardless of their condition of Expectancy.

When the final target subject had been scored the instructor thanked the observers for their cooperation and offered to send them, in a self-addressed envelope, a summary of the experiment and its outcome. This summary also, in fact, included a justification of the deceptions used. Finally, the observers were asked to complete a questionnaire, a copy of which is found in Appendix B. The purpose of certain crucial items, in this questionnaire, was to assess the observers' awareness of the real purpose of the experiment and to assess their expectations in relation to the different conditions of Expectancy. When the questionnaire was completed the experimenter signed the experimental credit cards and provided the envelopes mentioned in the instructions if the observers requested them. As the observers were leaving the experimenter expressed his appreciation of the way they had assisted in the investigation.

Method - Study II

Observers and Experimental Design

It had been intended to employ either the PA or the PP Condition

in this study, depending on the condition in which consensus was found. However, since consensus occurred in PA for nods and in PP for smiles in Study I, it seemed appropriate to use both PA and PP in the second study. Similarly, it had originally been intended to administer the N expectancy to all observers in this study but, since the consensus effect was confined mainly to the D expectancy in Study I, it was decided to use that expectancy in Study II. In the D Condition, it will be recalled, observers were told that the target subjects disliked the interviewer.

Sixty-four observers were selected from the same population as those employed in Study I. They were allocated randomly to the various experimental conditions, and took part in two experimental sessions, the second following a week after the first, as in Study I. In this study nine observers were eliminated through failure to show up for one of the two observational sessions, and a further observer because of a similar failure on the part of a confederate. These ten observers were replaced, as in the first study, by an equal number from the same population.

To actively manipulate consensus, observers participated as partners of one of two confederates who posed as fellow-observers and consistently scored either high or low. The two confederates each received the same training and participated to an equal extent in all experimental conditions.

Thirty-two observers, half of them in PA Condition and the other half in PP Condition, were paired with a High-Scoring Confederate (HSC). A further 32 observers, also equally divided between PA and PP Conditions, were paired with a Low-Scoring Confederate (LSC). Pairing conditions were administered as in Study I. That is, in PA Condition the confederate scored high or low during the practice session but did not participate in

the test session, while in PP Condition he scored high or low in both practice and test sessions.

Stimulus and Related Materials

The recordings of the five target subjects and of the instructions, which were used in Study I, were employed in the same manner in this study. The experiment was conducted in the same setting, using the same apparatus, as in Study I.

Procedure

Both experimental sessions followed the same procedure as in Study I, the experimenter and confederate being careful to minimize the confederate's contact with the observer immediately before and after the experimental sessions. Under the HSC Condition the confederate recorded eight smiles and twenty-four nods for each target subject. Under the LSC Condition the corresponding scores were two smiles and six nods. These frequencies were obtained by averaging the scores of two previous pairs of observers over the two test subjects and doubling and halving these averages for high and low scoring conditions respectively. In other words, these four earlier observers scored an average of four smiles and twelve nods for each of the two target subjects.

The confederate recorded half of the scores in the first five minutes of each trial and the remaining half in the second five minutes. He was instructed to relate his scoring as closely as possible, within the limitations imposed by having to record a predetermined score, to the actual responses of the target subjects. Otherwise the observers might have been aware that the confederate scored at a particular level but might not have been directly influenced by the confederate in scoring each individual response. Since consensus might operate through either,

or both, of these channels, i.e., by the observer being aware that his partner scored at a high or at a low level, or by the observer being aware that his partner did, or did not, score a particular response, it was necessary for the confederate to attempt to influence his partner through both of these channels.

CHAPTER IV

RESULTS

Study I

Questionnaire Data

The questionnaire was intended to provide information regarding the observers' awareness of the purpose of the experiment, and also to permit an assessment of the effectiveness of the Expectancy manipulations. The data obtained are presented here, in this study, rather than later, because they are highly relevant to the interpretation of the main findings for expectancy effects.

Purpose of Experiment. No observer in AA Condition of Pairing indicated any awareness of the true purpose of the experiment. In PA Condition one observer indicated awareness and two further observers gave ambiguous responses which may, or may not, have indicated awareness, while the responses of the remaining observers implied unawareness. All observers in PP Condition indicated unawareness, with the exception of one who was aware. In D Condition of Expectancy no observer was aware. Awareness, ambiguity, and unawareness were each distributed equally between the remaining two conditions of Expectancy. These data are presented in Table 14 in Appendix D. Thus, out of a total of 144 observers, two were aware of the purpose of the experiment, and two more may have been aware.

Expectancy for smiles. In L Condition of Expectancy 40 observers, out of a total of 48, responded as anticipated. That is, they expected the target subjects to increase or reduce their output of smiles in

positive relationship with the favourability of their attitude towards the interviewer. Six observers either stated that the subject's attitude to the interviewer would make no difference to his output of smiles, or failed to answer this part of the question. Two observers gave the opposite reponses to what had been anticipated. In other words, they expected the subjects to increase or reduce their output of smiles in negative relationship to the favourability of their attitude to the interviewer. Very similar results were obtained in N Condition. In D Condition only 32 observers responded as anticipated, 11 were uncertain or believed there would be no difference in the output of smiles, and five gave responses opposite to what had been anticipated. Between conditions of Pairing the variations in these responses were very small. In summary, out of a total of 144 observers, 111 responded as anticipated, 24 did not respond or expected no difference in output of smiles, and nine gave responses opposite to what had been anticipated. These data are presented in Appendix D, Table 15.

Expectancy for nods. In L Condition of Expectancy 13 observers responded as anticipated, 26 either stated that the subject's attitude to the interviewer would make no difference to his output of nods or failed to answer this part of the question, and nine observers gave the opposite responses to what had been anticipated. Almost identical results were obtained in N Condition. In D Condition 21 observers responded as anticipated, 16 either stated that the subject's attitude to the interviewer would make no difference to his output of nods or failed to answer this part of the question, while 11 observers gave the opposite responses to what had been anticipated. There were was little variation in these responses between conditions of Pairing. In summary, 47 observers, out

of a total of 144, responded as anticipated, 66 did not respond or expected no difference in output of nods, and 31 gave responses opposite to what had been anticipated. These data are presented in Table 1. They indicate that, for nods, the expectancy manipulation was not successful. It may therefore be anticipated that these expectancy manipulations would not affect the observers' scoring of nods.

Observer Expectancy Effect

All analyses were computed separately for smiles and for nods, the data employed being the scores of the observers for the responses of the two target subjects in the second observational session. The scores for the third target subject in the first observational session were used as the covariate in the analysis of covariance. The effects of two factors, each having three treatment conditions, were analyzed. The Expectancy factor was examined under three conditions of expectancy, L, N, and D, and the Pairing factor under three conditions of pairing, AA, PA, and PP. It was considered preferable to test the data for Expectancy effect using analysis of covariance, and to test the same data for Pairing effect using analysis of variance. The earlier study (Montgomery and Adair, 1971) had indicated considerable individual variation in the scoring of observers in practice trials, Analysis of covariance was therefore employed to equate the groups, statistically, in this respect. On the other hand, since the effects of pairing were already present in AA and PP Conditions when the baseline scores were obtained in the practice session, and since this effect was not manipulated within the experiment in these conditions, the use of analysis of covariance would have resulted in the influence of the different conditions of this factor being partialled out. To study the Pairing effect it was therefore necessary to use analysis of variance.

Table I

Direction of Observers' Expectancies for Nods

Variable	L			N			D			Totals		
	+[a]	0[b]	-[c]	+	0	-	+	0	-	+	0	-
AA	6	7	3	2	8	6	5	8	3	13	23	12
PA	3	10	3	5	8	3	7	4	5	15	22	11
PP	4	9	3	6	8	2	9	4	3	19	21	8
Totals	13	26	9	13	24	11	21	16	11	47	66	31

[a]Observers' expectancies were as anticipated.

[b]Observers either did not respond or indicated that the subjects' attitudes to the interviewer would not affect the output of nods.

[c]Observers' expectancies were opposite to those anticipated.

Smiles

Analysis of covariance. No significant differences or interactions were found, for smiles, in the overall analyses of either the summed scores or the separate scores for the two target subjects. Summaries of the analyses for the summed scores and for the separate scores for the two target subjects are presented in Tables 2 and 3, and in Tables 16 and 17 in Appendix D, respectively. They indicate that, while scoring seems to follow a similar pattern for both target subjects and the overall results are in the expected direction, the observer expectancy hypothesis was not supported. A casual examination of the cell means in Table 3, however, suggested that observers' scoring strategies differed somewhat between

Table 2

Analysis of Covariance of Observers Expectancy and Pairing Effects for Smiles

Source	df	MS	F
Expectancy (A)	2	27.406	0.38
Pairing (B)	2	17.674	0.25
A x B	4	144.984	2.01
Error	134	71.960	

Table 3

Adjusted Means for Observers' Expectancy and Pairing Effects for smiles

Variable	L	N	D	Total
AA	20.325	16.150	15.386	17.298
PA	18.137	21.467	15.707	18.440
PP	17.030	15.730	19.822	17.513
Total	18.505	17.749	16.996	

conditions of Pairing and that significant effects in one of these conditions may have been nullified, in the overall analysis, by a reversal in one of the other conditions. To investigate this possibility a separate analysis of covariance was computed for each condition of Pairing. The weakness of resorting to such post hoc analyses is recognized. However, the outcome, which is suggestive of a basis for further study, provides some justification for this procedure. It was found that the Expectancy effect was significant in AA Condition, ($F = 3.583$, $df = 2/44$, $p < .05$), and in PA Condition, ($F = 3.696$, $df = 2/44$, $p < .05$), but not in PP Condition, ($F = 1.225$, $df = 2/44$, $p > .05$). Summaries of these analyses are presented in Appendix D, Table 18. Analyses within pairing conditions indicated that AA observers in L Condition scored significantly more smiles than those in either N Condition, ($t = 2.2271$, $df = 44$, $p < .025$), or D Condition, ($t = 2.3958$, $df = 44$, $p < .025$), but N and D Conditions did not differ significantly, ($t = .1581$, $df = 44$, $p > .05$). For PA Pairing observers in N Condition scored significantly more smiles than those in D Condition, ($t = 2.7055$, $df = 44$, $p < .025$), but observers in L Condition did not differ significantly from those in N, ($t = 1.6068$, $df = 44$, $p > .05$), nor from those in D Condition, ($t = 1.1002$, $df = 44$, $p > .05$). In summary, while no overall Expectancy effects were found, the effect did occur in AA Condition, in which no partners were present in either session, some effect was found in PA Condition, in which observers were paired in the first session but not in the second, and there was no effect in PP Condition, in which observers were paired in both sessions. Such results as were significant were in the direction predicted in the Expectancy hypothesis. It thus appears that observers are more dependant on expectancy information when alone than when cues may be provided by a

partner, and will use that information to their advantage.

Observers' scores for the First and Second Periods were also tested separately, by analysis of covariance, in order to examine the influence of the duration of the observational trials on the Expectancy effects. For the First Period no significant main effects nor interactions were found. For the Second Period there was only a significant Expectancy x Pairing interaction. This interaction is difficult to interpret, being masked by the complexity of the various experimental conditions, but appears to be mainly due to a reversal of the Expectancy effect between AA and PP Conditions. Apart from this interaction, it appears that the operation of the Expectancy effect is not influenced by the duration of the trials. Summaries of these analyses are presented in Appendix D, Tables 19 and 20.

Analysis of variance. As explained earlier in this chapter, analysis of covariance is not a satisfactory test for Pairing effects and analysis of variance must be employed to examine this factor statistically. The main concern therefore, in this analysis, is with the effects of Pairing, and not with the effects of Expectancy, except in so far as the latter may interact with the former.

There were no significant main effects for Pairing, ($F = .96$, $df = 2/135$, $p > .05$), but there was a significant Expectancy x Pairing interaction ($F = 2.84$, $df = 4/135$, $p < .025$). This interaction, like that obtained in the analysis of covariance of the Second Period, above, seems to be due, in part, to a reversal of the Expectancy effect between AA and PP Conditions of Pairing, but may also be partly attributed to a reversal of the Pairing effect between L and D conditions of Expectancy. Summaries of these analyses are presented in Appendix D, Tables 21 and 22.

Nods

Analysis of covariance. No significant differences nor interactions were found for nods in the overall analysis of the summed scores for the two target subjects. This is as was anticipated on the basis of the failure of the Expectancy manipulation for this variable. Summaries of these analyses are presented in Tables 4 and 5. The question may arise as to whether significant results might have been obtained if the data had been analyzed separately for each of the two target subjects. Such analyses were in fact performed but the results were not significant for either target subject.

Analysis of variance. Here, as in the analysis of variance for smiles, the main concern was with the Pairing effects. No significant main effect nor interaction was found in the overall analysis of the summed scores for the two target subjects. Summaries of these analyses are presented in Appendix D, Tables 21 and 22. Inspection of the data, as well as the failure of the expectancy manipulation, suggested that there was no purpose in proceeding further with these analyses of the scores for nods.

Table 4

Analysis of Covariance of Observers' Expectancy and Pairing Effects for Nods

Source	df	MS	F
Expectancy (A)	2	222.480	0.46
Pairing (B)	2	225.775	0.46
A x B	4	39.583	0.08
Error	134	486.813	

Table 5

Adjusted Means for Observers' Expectancy and Pairing Effects for Nods

Variable	L	N	D	
AA	19.685	23.207	18.907	20.620
PA	21.631	26.164	25.365	24.380
PP	20.036	23.818	18.314	20.708
	20.456	24.397	20.856	

Observer Effects - Correlations

Observer effects were investigated by the comparison of agreement between the scores of paired observers with agreement between the scores of the same observers randomly matched, and by examining how these results were affected by the different conditions of Pairing. In AA Condition, of course, only agreement between the scores of randomly matched observers could be computed. The degree of agreement was assessed using Pearson product-moment correlations for the scores of the following sets of observers.

a) Randomly matched observers, within the same condition of Expectancy, who participated under AA Condition.

b) Members of pairs in PA Condition, each member of a pair being allocated to a different group for purposes of comparison.

c) Observers, who participated under PA Condition, randomly matched with others in the same conditions of Pairing and Expectancy.

d) Members of pairs in PP Condition, each member of a pair being allocated to a different group for purposes of comparison.

e) Observers who participated under PP Condition randomly matched with others in the same conditions of Pairing and Expectancy.

These correlational analyses were computed using the combined scores for the two target subjects in the second observational session within the one analysis, and were also computed separately for the scores of each of these subjects. The purpose of computing separate analyses for each of the target subjects was to eliminate the effect on the correlations of these scores having been recorded by the same observers and therefore being related. The magnitude of the correlation was taken

as a direct indication of the degree of agreement between observers, as has been the custom in studies employing paired observers. A test for the significance of a difference between two correlations was used to determine whether the various correlation coefficients that were to be compared differed significantly. Such a significant result was interpreted as indicating consensus between the observers.

Smiles

Pairing versus random matching within conditions of Pairing. The correlations for each of the pairings are presented in Table 6. The scores of randomly matched observers, participating under AA Condition, correlated at a level approximating zero (r = -.048, df = 22, p > .05). The correlations for partners and for randomly matched observers, under PA Condition, were also close to zero and the difference between them was not significant. Under PP Condition the correlation for partners was significant (r = .682, df = 22, p < .01), but that for randomly matched observers was negative and nonsignificant (r = -.105, df = 22, p > .05). The difference between these coefficients was significant (z = 4.408, p < .00004). The coefficient for AA Condition was not intermediate between those for the paired and randomly matched observers in PA Condition. This condition of intermediacy was fulfilled, however, in PP Condition, that is, in the Pairing condition in which consensus occurred.

When separate correlational analyses were computed for the scores for each target subject no major changes in the results were found. These analyses do provide the additional information that, under PP

Condition, the consensus effect was due mainly to the scores for the second target subject. For the first subject the correlations for paired and for randomly matched observers were .329 and -.029 respectively. These correlations were nonsignificant and the difference between them just failed to reach significance ($z = 1.72$, $p > .05$). The corresponding coefficients for the second target subject were .763 and -.135, the first being significant ($df = 22$, $p < .01$). The difference between them was also significant ($z = 5.363$, $p < .0000005$). Summaries of these analyses are presented in Appendix D, Table 23.

Table 6

Correlation Coefficients for Paired and Randomly Matched Observers, for Smiles

Variable	Paired	Randomized
AA		-.048
PA	.047	.195
PP	.682[a]	-.105*

*$p < .00004$, for difference between coefficients within row.

[a]$p < .005$, for correlation coefficient.

In summary, when the scores of the two target subjects were included in one correlational analysis the coefficients for the paired and randomized observers did not differ significantly under PA Condition but did differ significantly under PP Condition. The coefficient obtained

in AA Condition, and both coefficients obtained in PA Condition, were intermediate between those for paired and randomized observers in PP Condition. When analyses were computed separately for each target subject no important change was found in the pattern of results. The same significant differences were obtained as when the two target subjects were included in the one analysis, except that the difference between the coefficients for paired and randomized observers, under PP Condition, just failed to reach significance for the first subject.

Pairing versus random matching within conditions of Expectancy. Examination of the data raised the question of whether consensus was distributed equally over the different Expectancy conditions or was limited mainly to the D Condition. Since this issue was important in deciding which Expectancy condition to employ in Study II, the data for PA and PP Conditions were combined and an analysis for consensus performed, by means of the correlational procedures and tests for significance described above, for each condition of Expectancy.

For the purpose of this analysis the data was collapsed across PA and PP Conditions and an analysis, similar to that described above for Pairing versus random matching within conditions of Pairing, was performed for each condition of Expectancy. A summary of these analyses is presented in Appendix D, Table 24. In L Condition coefficients for paired observers (r = -.033) and for randomly matched observers (r = .091) were not significantly greater than zero, and the difference between them was not significant.

Similarly, in N Condition, correlation coefficients for paired observers (r = .135) and for randomly matched observers (r = .302) were not significant, nor were the differences between them significant. In D Condition, however, while the correlations for paired observers was significantly greater than zero (r = .699, df = 14, p < .005), the coefficient for randomly matched observers was negative (r = -.128). The difference between these two coefficients was significant (z = 3.699, p < .0002). Since this analysis was performed for the sole purpose of identifying the Expectancy conditions in which consensus operated, thereby providing procedural guidance for Study II, it is not related to any hypothesis. It was concluded that consensus does not operate, for smiles, in either L or N Condition but does operate in D Condition.

Mods

Paired versus random matching within conditions of Pairing. A summary of these analyses is presented in Table 7. The scores of randomly matched observers, participating under AA Condition, correlated at a level which was not significantly different from zero (r = -.149). Under PA Condition neither the coefficient for paired observers (r = .235) nor for randomly matched observers (r = -.110) reached significance but the difference between them was significant (z = 1.754), p < .05). Under PP Condition neither the coefficient for paired observers (r = .142) nor for randomly matched observers (r = .018) was significant, nor was the difference between them

significant (z = .588, p > .05). The correlation for randomly matched observers in AA Condition (r = -.149) was not intermediate between those for paired and for randomly matched observers in either PA or PP Conditions.

Table 7

Correlations Coefficients for Paired and Randomly Matched Observers for Nods

Variable	Paired	Randomized
AA		-.149
PA	.253	-.110*
PP	-.142	.018

*p < .05, for difference between coefficients within row.

When these data were analyzed separately for each target subject the pattern of results remained almost unchanged, as may be seen in the summary of analysis presented in Appendix D, Table 25. The only important difference was that the correlation coefficient for paired observers, under PA Condition, was now significant (r = .442, df = 22, p < .05). The differences between coefficients for paired and randomly matched observers, under PA Condition, remained significant. It may be noted that one of the significant differences mentioned here, as elsewhere in this study, includes a comparison between coefficients

which are not themselves significantly different from zero, i.e., for the first target subject in PA Condition. While some authorities contend that it is not legitimate to consider the significance between correlation coefficients when neither is itself significant, such a view is, at least, debatable.

In summary, when the scores for the two target subjects were combined in one correlational analysis the coefficients for paired and for randomly matched observers differed significantly under PA Condition but not under PP Condition. The same difference was significant for both subjects when analyses were computed separately for the two target subjects.

Pairing versus random matching within conditions of Expectancy. The rationale of this analysis was the same as that for the corresponding analysis for smiles. A summary of the analyses is presented in Appendix D, Table 26. It was found that in L Condition correlation coefficients for both paired observers (r = .177) and for randomly matched observers (r = -.222) were not significant (df = 22, p > .05), and the difference between them was also nonsignificant (z = 1.483, p > .05). Results were similar for N Condition where correlations for both paired observers (r = -.201) and

randomly matched observers (r = -.059) were negative. In D Condition, however, the coefficient for paired observers was significant (r = .531, df = 22, p < .025) while that for randomly matched observers was nonsignificant (r = -.023, df = 22, p > .05). A significant difference was found between these coefficients (z = 2.279, p < .003). It was concluded that consensus operates for nods, as for smiles, only in D Condition.

Observer Effects - Difference Scores

It will be realized that observers' scores might agree or differ not only as they vary in relation to each other, as measured by the correlational analyses, but also in the absolute values of the scores. Consensus between observers was therefore investigated by comparing the difference scores of paired observers with the difference scores of the same observers randomly matched. That is, the sum of the observer's scores for the two test trials was subtracted from the sum of the corresponding scores of the observer with whom he was paired or matched, the result being the difference score. The difference between the difference scores of paired observers and those or randomly matched observers was tested for significance, by means of the t test, in each of PA and PP Conditions of Pairing. A significantly lower difference score for paired observers than for randomly matched observers was

interpreted as indicating the operation of consensus.

Smiles

In PA Condition the mean difference score for paired observers was 10.50, and that for randomly matched observers was 10.92. The difference was not significant, (t = 0.202, df = 46, p > .05). A nonsignificant difference was also observed in PP Condition, in which the means for paired and for randomly matched observers were 7.50 and 10.50 respectively, (t = 1.337, df = 46, p > .05).

Nods

In PA Condition the mean difference score for paired observers was 20.71, and that for randomly matched observers was 26.71. Again, the difference was not significant, (t = 0.785, df = 46, p > .05). Likewise, a nonsignificant difference was obtained in PP Condition, where the means for paired and for randomly matched observers were 15.67 and 17.42 respectively, (t = 0.435, df = 46, p > .05).

In summary, all the differences, for both smiles and nods, were in the expected direction. That is, difference scores for randomly matched observers exceeded those for paired observers. None of the differences, however, was significant.

Summary of Results - Study I

No overall effects were found for observer expectancy, but the effect was found for smiles when observers scored alone. The absence of this effect for nods may be due to the failure of the expectancy manipulation for this variable.

The observer consensus effect occurred, for smiles, when observers were paired in both practice and test sessions. The effect occurred, for nods, when observers were paired in the practice session but scored

alone in the test session. The reliability coefficients for observers scoring alone in both sessions were close to zero for both smiles and nods. When the data were analyzed separately for each target subject the results were very similar to those obtained using the combined scores for these subjects.

Study II

In this study the scores of observers in HSC Condition were compared with the scores of those in LSC Condition. All these observers scored under D Condition of Expectancy, i.e., they were told that the target subjects disliked the interviewer. The analyses were performed separately for smiles and nods. Overall tests by analysis of variance were carried out both for the summed scores of the two target subjects and for each of these subjects separately. The effects were also analyzed by *t* tests for each of PA and PP conditions of Pairing, these tests, like the analyses of variance, being computed for the summed scores for the target subjects and for each of these subjects separately. To eliminate the possibility of unintended confederate effects the scores of observers who were paired with one confederate were compared with the scores of those who were paired with the other and tested for significant differences, also by means of the *t* test, at each confederate level of scoring.

Smiles

Overall effects of confederates' levels of scoring. In the analysis of variance using the summed scores for the two target subjects the scores of HSC observers were significantly greater than those of LSC observers, ($F = 5.08$, $df = 1/60$, $p < .05$). The two conditions of Pairing did not

differ significantly from each other, (F = 3.17, df = 1/60, p > .05), nor was there a significant interaction, (F = 0.17, df = 1/60, p > .05). These results are presented in Tables 8 and 9. When the data were analyzed separately for the two target subjects it was found that while scores of HSC observers were significantly greater than those of LSC observers for the first subject, (F = 5.30, df = 1/60, p < .025), the corresponding scores for the second subject were not significantly different, (F = 2.79, df = 1/60, p > .05), although they did differ in the predicted direction. There was also a significant difference for Pairing for the first target subject, (F = 4.66, df = 1/60, p < .05), observers in PA Condition recording higher scores than those in PP Condition. These analyses are presented in Tables 10 and 11.

Effects of confederates' levels of scoring in each condition of Pairing. When the t test was applied to PA Condition no significant difference was found between the scores of observers in HSC and LSC Conditions, (t = 1.087, df = 30, p > .05), but the corresponding difference was significant in PP Condition, (t = 2.504, df = 30, p < .01). All t tests throughout this study, except those for unintended confederate effects, are one-tailed. Separate analyses for each of the target subjects indicated that, in PA Condition, results were not significant either for the first target subject, (t = 1.231, df = 30, p > .05), or for the second target subject, (t = 0.631, df = 30, p > .05). In PP Condition, however, significant differences were obtained for both the first target subject, (t = 2.540, df = 30, p < .01), and for the second target subject, (t = 2.144, p < .025).

Table 8

Analysis of Variance for Smiles, with Observers Paired under Two Conditions with Confederates Scoring at High and Low Levels

Source	df	MS	F
Level (A)	1	360.9983	5.08*
Pairing (B)	1	224.0995	3.17
A x B	1	12.2500	0.17
Error	60	71.0582	

*$p < .05$.

Table 9

Mean Scores for Smiles, with Observers Paired under Two Conditions with Confederates Scoring at High and Low Levels

Conditions of Pairing	High	Low	Total
PA	18.62	14.75	16.69
PP	15.75	10.13	12.95
Total	17.19	12.44	

Confederate effects. The mean score of observers paired with Confederate A did not differ significantly from that of observers paired with Confederate B, either in HSC Condition, (t = 0.136, df = 30, p > .05), or in LSC Condition, (t = 1.360, df = 30, p > .05). The mean scores, for both smiles and nods, are presented in Appendix D, Table 29.

In summary, the confederates' levels of scoring did not influence observers' scores significantly in PA Condition. They did, however, have a significant influence in PP Condition but it was confined to the first target subject. There were no significant differences between the scores of observers paired with different confederates.

Nods

Overall effects of confederates' levels of scoring. In the analysis of variance, using the summed scores for the two target subjects, the scores of HSC observers were significantly greater than those of LSC observers, (F = 21.90, df = 1/60, p < .001). Results for the two conditions of Pairing did not differ significantly, (F = 2.03, df = 1/60, p > .05), nor was there a significant interaction, (F= 0.12, df = 1/60, p > .05). These results are presented in Tables 12 and 13. The analyses of the data, computed separately for each of the two target subjects, indicated that scores of HSC observers were significantly greater than those of LSC observers both for the first subject, (F = 20.34, df = 1/60, p < .001), and for the second subject, (F = 20.18, df = 1/60, p < .001). These results are presented in Tables 27 and 28.

Effects of confederates' levels of scoring in each condition of Pairing. The application of the t test to the data for PA Condition indicated that the scores of HSC observers were significantly greater than those of LSC observers, (t = 3.981, df = 30, p < .0005). A

Table 10

Analysis of Variance for Smiles, First and Second Target Subjects

Source		First Subject		Second Subject	
	df	MS	F	MS	F
Level (A)	1	143.9996	5.30**	48.9999	2.97
Pairing (B)	1	126.5621	4.66*	14.0625	0.85
A x B	1	0.5621	0.02	7.5625	0.46
Error	60	27.1812		16.5187	

*p < .05

**p < .025

Table 11

Means for Smiles, for First and Second Target Subjects

Conditions	Subject 1			Subject 2		
of Pairing	High	Low	Totals	High	Low	Total
PA	13.56	10.75	12.16	5.06	4.00	4.53
PP	10.94	7.75	9.34	4.81	2.38	3.59
Total	12.25	9.25		4.94	3.19	

Table 12

Analysis of Variance for Nods, with Observers Paired under Two Conditions with Confederates Scoring at High and Low Levels

Source	df	MS	F
Level (A)	1	7876.5469	21.90*
Pairing (B)	1	42.2494	0.12
A x B	1	729.0005	2.03
Error	60	359.6094	

*$p < .001$.

Table 13

Mean Scores for Nods, with Observers Paired under Two Conditions with Confederates Scoring at High and Low Levels

Conditions of Pairing	High	Low	Total
PA	42.44	13.50	27.97
PP	34.06	18.63	26.34
Total	38.25	16.06	

significant difference was also found for the corresponding groups of observers in PP Condition, (t = 2.536, df = 30, p < .01). Separate analyses for each of the target subjects indicated that in PA Condition results were significant both for the first target subject, (t = 3.417, df = 30, p < .005) and for the second target subject, (t = 4.192, df = 30, p < .0005). Significant differences were also obtained in PP Condition for both the first target subject, (t = 2.964, df = 30, p < .005), and for the second target subject, (t = 2.084, df = 30, p < .025).

Confederate effects. The mean score of observers paired with Confederate A did not differ significantly from that of observers paired with Confederate B, either in HSC Condition, (t = 1.381, df = 30, p > .05), or in LSC Condition, (t = 0.991, df = 30, p > .05). The mean scores are presented in Appendix D, Table 29.

In summary, the confederates' levels of scoring influenced the scores of observers significantly in both PA and PP Conditions. There were no significant differences between the scores of observers paired with different confederates.

Questionnaire Data

The questionnaire employed in the first study was also used, in the same manner, in this experiment. Its main purpose was to provide information regarding the observers' awareness of the true purpose of the study. It should be noted that the purpose of this study was to investigate the effect of the confederates' different levels of scoring on the scoring of the observers, not to investigate the effects of Expectancy and the influence of one observer on another as in Study I. A secondary purpose of the questionnaire was to provide further data on the observers' expectancies, for comparison with similar data derived from the first

study.

Purpose of the experiment. No observer, out of the 64 who participated, indicated any awareness of the true purpose of this experiment. Four of the observers, however, suspected that the purpose of the experiment was the study of the effects of observers' expectancies, as in Study I.

Expectancy for smiles. Fifty-two of the 64 observers responded as anticipated, while seven either stated that the subject's attitude to the interviewer would make no difference to his output of smiles or failed to answer this part of the question, and five gave the opposite responses to what had been anticipated.

Expectancy for nods. Twenty-two observers responded as anticipated, 26 either stated that the subject's attitude would make no difference or did not answer, and 16 gave the opposite responses to what had been anticipated. It will be recalled that all observers in this study participated under D Condition of Expectancy.

In summary, expectancies were in the anticipated direction for smiles but not for nods, a result similar to the overall pattern for smiles and nods in the first study.

Summary of Results - Study II

The manipulation of observers' scoring, by pairing them with confederates scoring at predetermined high and low levels, was effective, for both smiles and nods. Observers paired with high-scoring confederates scored significantly more responses than those paired with low-scoring confederates. For smiles the effect was found only for observers scoring under PP Condition. The effect was found for nods in both PA and PP Conditions but was stronger in the former. Thus the results closely parallel those found for consensus in the first study.

CHAPTER V

DISCUSSION

Observer Expectancy Effects

In the first study one of the factors explored was the contaminating effect of observers' expectancies upon their scoring of subjects' non-verbal responses. This effect was found when observers scored alone in both practice and test sessions. Observers who had been told that the subjects liked their interviewer scored more smiles than either those who had been told that the subjects were neutrally disposed towards the interviewer or had been told that they disliked him. The effect was found, to a lesser extent, when the observers were paired in the practice session and scored alone in the test session. In this situation fewer smiles were scored for subjects who supposedly disliked the interviewer than for subjects presented as neutral. A somewhat complex interaction occurred between Expectancy and Pairing Conditions. In one condition of Pairing, i.e., the PA, Expectancy effects were as predicted while in the other, the PP Condition, they were in the opposite direction, with a similar change of direction occurring for Pairing Effects from one condition of Expectancy to another. No significant effects nor interactions were found for nods. A separate analyses of both smiles and nods for each of the two target subjects was undertaken in order to examine any difference between the effects for each of these subjects. These analyses yielded results similar to those for the combinations for both target subjects. It may therefore be concluded that expectancies did not have

any serious influence on the observers' scoring, except for the scoring of smiles by unpaired observers.

The complete absence of any expectancy effect for nods is a logical outcome of the failure of the expectancy manipulations for this variable. This failure presents a problem of some difficulty since the same manipulation in the earlier study (Montgomery and Adair, 1971) was approximately as effective as that for smiles, as indicated by the questionnaire responses. The most likely explanation that can be suggested is that the investigator's expectancy was communicated non-verbally to the observers in the earlier experiment while in the present study the relatively uninvolved instructor communicated little or no expectancy and the observers, consequently, failed to form expectancies.

The fact that no expectancy effect was found for smiles, when observers scored in pairs, can not be attributed to any similar failure of the expectancy manipulation. It appears, rather, that when more than one source of contamination is present the effect of each is attenuated. That is, paired observers may have taken advantage of scoring cues transmitted by their partners and, as a result, felt less need for cues provided by expectancies. On the other hand, observers who had no partners to provide scoring cues may have availed themselves to a greater extent of the expectancy cues.

What this investigation demonstrated regarding the expectancy effect should be clarified. No attempt was made to demonstrate an expectancy effect created by nonverbal cues from the experimenter or from the target subjects. However, the investigation demonstrated that an expectancy created by information contained in the instructions influenced observers who scored alone in their recording of smiles. The fact that contamination

occurred in this particular situation should alert investigators to suspect its occurrence at least when these conditions are again present. Indeed, to the extent that such contamination has been demonstrated, it is a very real consideration that expectancy effects may occur in any study in which expectancies may be self- or procedurally-generated.

Observer Consensus Effects

Consensus, the process by which paired observers approach scoring agreement by means of an exchange of nonverbal scoring cues, was demonstrated in the first study. It occurred for smiles when observers were paired in both practice and test sessions and for nods when observers were paired in the practice session and scored alone in the test session. These results were obtained when the scores of the two target subjects were combined in one analysis as well as when they were computed separately for each of these subjects, with the exception of the scores for smiles for the first target subject for which the effect just failed to reach significance.

In these analyses the only significant coefficients were for smiles when observers were paired in both sessions and for nods when observers were paired in the practice session and scored alone in the test session. All coefficients for randomly matched observers were approximately zero or were negative. This was so irrespective of whether the scores of the two target subjects were combined or were

analyzed separately. In fact, scores for the first target subject did not reach significance under any condition, either when the observers were paired or were randomly matched. Nevertheless a consensus effect did occur for the first target subject for nods and, as mentioned above, the effect was just short of significance for smiles.

Further statistical testing was interpreted as indicating that for both smiles and nods consensus was confined to observers who had been told that the test subjects disliked the interviewer. Only speculative explanations can be offered for this unexpected finding. It may be that a person's expectations for another's behavior are less well-defined when he knows that the other dislikes someone else than in other circumstances. For example, when one person likes another he will usually tend to express his liking. However, in the case of disliking, he may either express his dislike or attempt to conceal it to avoid drawing hostile retaliation. Because of this uncertainty observers with an expectancy of dislike may have given greater attention to cues from their partners and thereby attained greater consensus. Tagiuri, Bruner, and Blake (1958) have discussed a somewhat similar result. They suggest that the behavior of persons who reject others may be particularly difficult to judge, partly because the rejection is usually concealed by politeness and partly because rejection normally terminates a

relationship, thereby limiting the opportunity to test rejection cues. Applying such reasoning to the present investigation, it is possible that observers were uncertain how much politeness to attribute to subjects who were supposed to dislike the interviewer. For example, they may have been in doubt whether to interpret a subject's response as a smile or as a politely disguised sneer. If indeed, as suggested, a situation of rejection reduced opportunities to test cues, the observers may have assumed that their lack of training in interpreting such cues left them poorly equipped for their task. Whatever the explanation, it is clear that this tendency of the consensus effect to occur most readily in one particular condition of expectancy is due to a difference in the subjective state of the observers, and not to an objective difference in the difficulty of scoring the target subjects, since all observers viewed identical target subjects.

While the overall consensus effect operated in accordance with predictions for smiles when observers were paired throughout both sessions, a surprising finding is that this effect operated even more strongly for nods when the observers scored as pairs in the practice session and scored alone a week later. This result suggests that these observers, focusing on a small video monitor and concentrating on the task of scoring responses, were able, incidentally, to obtain scoring cues from their companions during the practice session which considerably modified their response definitions in the direction of consensus which took effect during the test trials, for these observers. It may also be concluded that this change in response definition was of an enduring nature to the extent that, when these observers scored alone a week later, the consensus effect was stronger than for observers who then had direct scoring cues

available from their partners. It seems appropriate to think of this influence process as being initially unintended. These observers did not know during the first session that they would have no scoring partners in the second session and therefore had no reason to anticipate that direct scoring cues would not be available in the later session, rendering consensus possible only through agreement on response definitions.

The question then arises as to why consensus was demonstrated to a greater extent by observers scoring nods, when they scored as partners in the practice session and scored alone in the test session, than when they scored as partners in both sessions. No unequivocal explanation can be offered. It may be speculated that when observers realized, immediately after the first session, that they would score alone in the second session they felt some concern regarding the quality of their performance in scoring nods alone. A few comments of the observers, in questionnaire responses or in conversation with confederates, indicate that observers believed nods were more difficult to score than smiles. It would therefore seem that a process analogous to reminiscence may have occurred. That is, in his concern for reaching a satisfactory standard in scoring nods, the observer mentally rehearsed what occurred during the first session, including the cues provided by his partner. Since he then knew that these were the only consensus cues that would be available, and since he had ample time for this kind of rehearsal in the week between the two observational sessions, it is likely that he would attach more weight to these response definition cues than the observer who had a partner in the second session would attach to the directly available scoring cues as they occurred within the limited time of the observational session. As a consequence, the former observer was apparently more strongly

influenced by response definition consensus than the latter was by direct scoring cue consensus.

Observers who scored alone in both practice and test sessions also provided some particularly interesting results. These observers, by the nature of their experimental condition, were not exposed to any consensus influence. In that respect they represented an optimal scoring situation, and appear to permit a true assessment of observer performance. This was disappointing in that all correlation coefficients for this condition were negative, with the exception of that for the first target subject for smiles, which just exceeded zero. One may apparently conclude that, when not masked by artifact, the scoring performance of observers is extremely poor as measured by the accepted testing procedure. It may be suggested that observers might have improved with further practice. However, one may doubt that they would have attained a satisfactory standard of performance after any experimentally feasible degree of practice when it is considered that the reliability coefficients of these observers, in the testing session, were generally negative and that this was after they had already spent thirty minutes practicing in the earlier session.

Thus, the use of the correlation coefficient in this study as a measure of observers' scoring reliability has been shown to be very misleading, particularly when it is considered that all coefficients for randomly matched observers were approximately zero or were negative. The implications for the use of such reliability coefficients in the traditional manner, as a measure of observer agreement, are very serious

indeed.

Consensus was also investigated by a comparison of the difference scores of paired and randomly matched observers. This additional test for consensus was employed to provide a measure of absolute differences between the observers' scores, as distinct from the measure of relative difference provided by the correlational analyses. It was anticipated that consensus between paired observers would result in their difference scores being lower than those of the randomly matched observers. Although the scores differed in the expected direction for both smiles and nods none of the differences was significant.

It might have been expected from the demonstration of consensus by the correlational analyses that similar results would have been obtained by the analysis of difference scores. The failure of the latter analysis to show significant differences between the scores of paired and of randomly matched observers is open to at least two interpretations. The statistical tests employed were necessarily different in the two analyses and it is possible that one was less sensitive than the other to differences in observer agreement. It is also possible that difference scores measure a form of observer agreement which is largely independent of that assessed by the correlational analysis. Indeed the purpose in using two different measures of agreement was to test this possibility. Observers' scores may fluctuate in some degree of conformity with those of their partners, as the correlational analyses indicate, without reducing the difference between the absolute values of these scores. In terms of this investigation, paired observers may mutually agree, through consensus, to raise or lower both their levels of scoring without affecting the number of units by which their scores differ. This would result in the differences

which were found by correlational analyses between paired and randomly matched observers and would not require any corresponding lack of agreement for the difference scores.

The foregoing discussion of contamination by consensus indicates that, while this effect is strongly operative in certain circumstances, it has not been demonstrated in all situations in which observers scored simultaneously. Unfortunately this does not imply that there are situations in which satisfactory observer performance may be assumed. In this study the occurrence of consensus appears to depend on an interaction of several factors; pairing conditions, expectancy conditions, response category, and the target subject. It would therefore seem hazardous to generalize from the negative findings of this study to other experiments in which different conditions are employed. That is, the complexity of the interactions necessitates testing for consensus in each particular experimental situation.

However, consensus was clearly demonstrated, for both response categories, in one or the other condition of pairing, indicating that this is a likely source of contamination whenever observers score simultaneously. From the results for nods it may also be concluded that even if observers score alone in test trials a strong consensus effect may result if they have previously scored together in practice.

Manipulated Consensus Effect

The variable investigated in the second study is not, strictly speaking, consensus, a term that implies a mutual approach towards agreement. The confederates, by the nature of their task, were not at liberty to adjust their scoring in the direction of agreement. Only the observer was free to adjust his scoring towards agreement with his

partner. That is, from the observer's point of view the process is the same in both studies.

The results of this second study, in which observers were paired with confederates scoring consistently at either a high or a low level, strongly supported those of the correlational analyses for consensus in the first study. Analyses of the summed scores for the two target subjects indicated that observers paired with a high-scoring confederate scored significantly more smiles and nods than did those paired with a low-scoring confederate. Further analyses indicated that this effect occurred, for smiles, when observers were paired in both practice and test sessions. Under these circumstances the effect was also demonstrated for nods. It was demonstrated even more strongly for nods, however, when observers were paired in the practice session and scored alone in the test session. Results similar to these were obtained for both smiles and nods when the analyses were computed separately for each target subject.

The results of the second study strengthen considerably the conclusions which may be based on the findings in the first. Not only was the influence of the confederate upon his companion demonstrated, but the pairing conditions in which this influence was found followed a pattern very similar to that in the first study. What is more important, the effect was demonstrated in the second study by experimental manipulation. The results of the first study were open to several alternative explanations. For example, differences between correlations of random matchings and original partners could have been due to the possible differential influence of the experimenter on pairs of observers, or a similar influence arising from the weather or other varying environmental conditions. These potential influences, involving common intra-pair

experiences not shared by other pairs of observers, could have inflated the correlations between paired observers and reduced the correlations between randomly matched observers. No such influences, however, can account for the correspondence between the scoring levels of observers and those of the confederates with whom they were paired. It therefore appears likely that observers scoring simultaneously under certain common conditions may seriously contaminate each other's scoring outputs by their mutual influence, even when precautions are taken to prevent this influence and when their simultaneous scoring is limited to the practice session.

Conclusions

On the basis of the results of these two studies one may speculate on the nature of the effect of consensus on the individual observer's scoring performance. Could the information conveyed by the scoring cues actually enhance the standards of scoring? For example, might the less objective observer be guided by the cues from his more objective companion, rather than vice versa, and thus improve his performance? Studies of reflective and impulsive responders suggest that this is not so. Kagan, Rosman, Day, Albert, and Phillips (1964), and Siegelman (1969), have shown among child subjects that impulsive responders, i.e. those responding more quickly, are less accurate than reflective responders who are slower in indicating their responses. This suggests that the first observer to score a response may be less accurate than his partner and will therefore have an unfavourable influence if the latter responds to the scoring cues he provides. Moreover, if consensus enhanced the observers' scoring performance, one would expect to find higher correlation coefficients than those actually obtained for randomly matched observers.

In the second study the observers were not deterred from accepting scoring cues from the confederates, in spite of the fact that the latter's level of scoring was determined by the investigator and not by the responses of the target subjects. It therefore seems evident that observers do not discriminate in the quality of the scoring cues which they allow to influence their own responses.

In assessing the implications of this discussion of consensus one further conclusion seems justified. The use of the reliability coefficient as a measure of the scoring performance of paired observers is a very questionable practice unless the most stringent precautions are taken to prevent consensus occurring, precautions which, as yet, can not be clearly defined.

While most of the positive findings are well supported in this investigation it may be asked whether they can be generalized to other observational situations or apply only to the experimental situation employed here? It should be remembered that the observational situation was deliberately designed as typical of those commonly found in investigations in social psychology. The conditions were not selected because of prior knowledge that they were more likely than others to produce the expected effects. In fact, by the use of videotape recordings, by employing a naive instructor, by placing a partition between the observers, by seating the observers at a distance from each other, and by instructions regarding independent scoring, the investigator went considerably beyond the precautions usually taken to reduce contaminating influences. It might also be expected that observers scoring a comparatively small number of target subjects, as in these studies, would have less exposure to contamination than would one pair of observers scoring a much larger number of

target subjects, which is the common procedure.

This investigation has demonstrated some of the pitfalls in the use of the observer technique. Unfortunately it is difficult to suggest remedies that would ensure satisfactory observer performance. The problem, clearly, is to prevent the occurrence of both expectancy and consensus effects. Pairing the observers in both practice and test sessions appears to contribute to the elimination of the expectancy effect. The consensus effect may obviously be removed by having the observers score alone in both practice and test sessions. However, the results of the first study indicate that, under these circumstances, significant reliability coefficients may be difficult to obtain. In addition, an investigator finds himself in a dilemma. If observers score alone consensus is eliminated, but this is the situation in which expectancy effects are most likely to occur. The problems of using the observer technique thus appear to be formidable and, for some of them, no satisfactory solution can be offered.

In summary, this investigation has shown that observers' expectancies can influence their scoring when they score alone. It has also shown that observer consensus occurs, under several combinations of conditions, when observers score in pairs, and that this effect may be found even when observers score alone if they have previously practiced together. The onus is therefore on the investigator who uses the observer technique to demonstrate that expectancy and consensus effects do not occur in his particular observational situation, or, at least, to point out that these effects may have influenced his results.

SUMMARY

This investigation explored the contaminating influences which appear to affect observers used to record the nonverbal responses of subjects in psychological experiments. The problems which may arise are serious since observers are widely used for this purpose, and are sometimes the only means of obtaining the data, and because the danger of observers being susceptible to such influences has been largely ignored.

An historical survey of the use of the observer technique reveals that investigators have been generally more concerned with the technical aspects of the problem than with the social psychology of the observer. A few studies have shown that observers' expectancies may differentially influence their scoring. Apart from rare theoretical speculations it has been assumed that the influence of paired observers upon each other is not a matter for concern. It has also been assumed that high observer agreement, as measured by reliability coefficients, is an indication of objective scoring. Consideration has not been given to the possibility that it may be due to inter-observer influence. Two aspects of observer contamination were therefore studied in this investigation; the effects of observers' expectancies on their scoring and of inter-observer influences.

Observers participated in a practice session followed by a testing session a week later. They scored the smiles and nods of several subjects presented on identical videotape recordings to all observers. Expectancies were manipulated by telling different groups of observers

that the subjects viewed in the test session liked or disliked the person interviewing them, or were neutrally disposed towards him. Results indicated that the observers' expectancies influenced their scoring when they scored alone. The observers were examined by three different procedures to determine inter-observer influences. Some observers scored alone in both practice and test sessions. Others were paired in the first session and scored alone in the second. Observers in a third group were paired in both sessions. The influence of paired observers upon each other was assessed by comparing the level of scoring agreement between partners with that of the same sets of observers randomly matched and with that of observers who had scored alone. Inter-observer influence was demonstrated when observers scored smiles, paired in both practice and test sessions, and when they scored nods, paired in the practice session but alone in the test session, the strength of these effects being related to the conditions of observer expectancy. In a second study an attempt was made to manipulate inter-observer scoring consensus by pairing the observers with confederates who scored at a consistently high or low level. The effect was demonstrated for smiles when the observers were paired with the confederates in both practice and test sessions. It was demonstrated for nods when observers were paired with confederates in both sessions and also when they were paired only in the practice session, the effect being stronger in the latter condition. Thus, the results of the second study not only strengthened the conclusions that may be based on those of the first, but indicated that inter-observer consensus can be demonstrated by experimental manipulation.

It was concluded that the observer technique is not free from contaminating influences, as had generally been supposed. Specifically, contamination due to observer expectancies and to observer consensus has

been demonstrated. These findings raise serious questions regarding studies in which observers are employed. Not only must precautions be taken in such studies to eliminate these subtle influences, but much intensive investigation is required.

References

Allport, F.H. Social psychology. Cambridge, Mass.; Riverside Press, 1924

Asch, S.E. Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), Groups, leadership and men. Pittsburgh: Carnegie Press, 1951, Pp. 177-190.

Bernhardt, K.S., Millichamp, D.A., Charles, M.W., & McFarland, M.P. An analysis of the social contacts of preschool children with the aid of motion pictures. University of Toronto studies: Child Development series. No. 10, 1937.

Campbell, D.T. Systematic error on the part of human links in communication systems. Information and Control, 1958, 1, 334-369.

Chapple, E.D. The interaction chronograph: Its evolution and present application. Personnel, 1949, 25, 295-307.

Cordaro, L., & Ison, J.R. Psychology of the scientist: X. Observer bias in classical conditioning of the planarian. Psychological Reports, 1963, 13, 787-789.

Dudycha, A.L., & Naylor, J.C. The effects of variations in the cue R matrix upon the obtained policy equation of judges. Educational and Psychological Measurement, 1966, 26, 583-603.

Exline, R.V. Explorations in the process of person perception: Visual interaction in relation to competition, sex, and need for affiliation. Journal of Personality, 1963, 31, 1-20.

Haith, M.M. A semiautomatic procedure for measuring changes in position. Journal of Experimental Child Psychology, 1966, 3, 289-295.

Heyns, R.W., & Lippitt, R. Systematic observation techniques. In G. Lindsey (Ed.), Handbook of social psychology. Vol. 1. Cambridge, Mass.: Addison-Wesley, 1954, Pp. 370-404.

Johnson, D.M., & Vidulich, R.N. Experimental manipulation of the halo effect. Journal of Applied Psychology, 1956, 40, 130-134.

Kagan, J., Day, D., Albert, J., & Phillips, W. Information processing in the child: Significance of analytic and reflective attitudes. Psychological Monographs, 1964, 78 (1, Whole No. 578).

Ling, D., Ling, A.H., & Doehring, D.G. Stimulus, response, and observer variables in the auditory screening of newborn infants. Journal of Speech and Hearing Research, 1970, 13 (1), 9-18.

Mehrabian, A., & Williams, M. Nonverbal contaminants of perceived and intended persuasiveness. Journal of Personality and Social Psychology, 1969, 13, 37-58.

Montgomery, D., & Adair, J.G. The social psychology of observers in the study of nonverbal behaviour. Unpublished manuscript, 1971.

Orne, M.T., & Scheibe, K.E. The contribution of nondeprivation factors in the production of sensory deprivation effects. The psychology of the "panic button". Journal of Abnormal and Social Psychology, 1964, 65, 3-12.

Pawlicki, R. Behaviour-therapy research with children: A critical review. Canadian Journal of Behavioural Science, 1970, 2 (3), 163-173.

Rosenfeld, R. Instrumental affiliative functions of facial and gestural expressions. Journal of Personality and Social Psychology, 1966, 4, 65-72.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. Rosnow (Eds.) Artifact in Behavioral Research. New York: Academic Press, 1969, 186-194.

Rosenthal, R., & Halas, E.S. Experimenter effect in the study of invertibrate behavior. Psychological Reports, 1962, 11, 251-256.

Schwartz, M.S., & Schwartz, C.G. Problems in participant observation. American Journal of Sociology, 1954, 60, 343-353.

Sherif, M. The psychology of social norms. New York: Harper & Row, 1936

Siegelman, E. Reflective and impulsive observing behavior. Child Development, 1969, 40, 1213-1222.

Smith, H.C. Sensitivity to people. New York: McGraw-Hill, 1966.

Syz, H.C. Observations on the unreliability of subjective reports of emotional reactions. British Journal of Psychology, 1926, 17, 119-126.

Thorndike, E.L. A constant error in psychological ratings. Journal of Applied Psychology, 1920, 4, 25-29.

Vidich, A.J. Participant observation and the collection and interpretation of data. American Journal of Sociology, 1954, 60, 354-360.

APPENDIX A

INSTRUCTIONS PRESENTED TO OBSERVERS ON VIDEOTAPE RECORDINGS

Study I

First Observational Session

The following instructions were read to observers in PP Condition of Pairing

I would like to welcome you to this experiment. You may be interested to know that I am not going to use you as subjects. The subjects have come and gone, and I have recorded their behaviour on videotape. All I want you to do is to watch them on this screen and to indicate, on a piece of paper, when they smile and when they nod. A number of others will do exactly the same task. I hope to get accurate data by taking the average of all your scores. That is, I am using the subject pool not to get subjects but to get observers. Of course, you will receive experimental credit for participation as observers rather than as subjects. I hope you will find the task interesting and that you will perform it as carefully as you can. I am going to give the instructions over the television set. With the subjects already recorded on videotape it is easier for me to appear on the tape too.

Now, I want you to listen carefully. When I have finished speaking you will see the movies we took of the subjects who were participating in the experiment. A hidden camera was used and the subject was unaware, at the time, that a recording was being made. Later he was shown the film and gave his consent to its use for experimental purposes. You will see the subject speaking to an interviewer who does not appear in the picture.

There were three different kinds of subjects. Before the interview they had each been given the task of working with the interviewer in the construction of a simple crossword puzzle. This is a task that drives most people either to enthusiastic co-operation or to bitter disagreement. At the end of the task the subject completed a questionnaire indicating how much he liked or disliked his co-worker. Afterwards we selected those who were neutral, that is, who neither liked him nor disliked him. What I am interested in is the effect of these different conditions on the frequency of the subjects' smiles and nods.

Your task will be to record every smile and nod you see the subject using. You will see each subject for ten minutes. Scoring begins as soon as the subject appears on the screen and continues until he disappears.

You will score on these rolls of paper tape (Instructor held up a roll of paper tape in its dispenser). First, you will put your name on the tape. Then, above your name, you will put the number of the subject you are going to score. I will tell you the number of the subject, and it will also appear on the screen. When the subject appears you will score by putting an "X" on the paper when he smiles, and by putting an "O" when he nods. Do take care that the Xs and the Os don't get on top of each other. Score them in a column, and keep pulling the paper towards you as you need it.

I need hardly define a smile, since everybody knows one when he sees it. For the purposes of this experiment a nod is defined as "a two-directional movement of the head on the vertical plane." That is, there are two movements in a nod (Instructor illustrated with his own head)--up and down, or down and up. An unbroken series of repeated nods is scored as one nod. By "an unbroken series of repeated nods" it is meant that the subject does not interrupt his nodding either by pausing or by alternative behaviour. He just keeps nodding (Instructor illustrated an unbroken series of repeated nods)--like that. You score only one "O" for all those nods. But, if he stops and starts again, then you put another "O".

Using these paper rolls you will find it quite easy, after a little practice, to score without taking your eyes off the subject. Do not look down at the paper while you are recording on it or moving it. At all times focus on the face of the subject. As you score keep pulling the paper towards you, so that the record is in a vertical sequence. If you should forget whether to use an X or an O you will see which is which on the sign just below the television screen.

I am also interested in changes in the frequency of smiles and nods over the ten-minute interview. Does the subject smile, or nod, more, or less, as time goes on? In order that time may be indicated in the record my assistant will ring a bell at the half-way mark of each interview, that is, five minutes after it began. (The sound of the bell was heard on the recording). When you hear the bell draw a line across the strip of paper and continue recording beyond the line. At the end of the ten minutes, when the subject has disappeared from the screen, draw another line, beyond the last score, and tear the paper off beyond that line. After that you will have a short break to relax, but don't go away. Stay seated in your chair. After the break the same procedure will be followed in scoring the second subject.

I'm sure you realize that you will need some practice to be able to score accurately, so we are going to begin with a few subjects who are not being used in the experiment.

Before we start, I want you to notice that the position of the front legs of your chair has been marked on the floor. This is to make sure that everybody is at the same distance from the screen and at the same angle. Would you check that your chair legs are at the marks? . . . The partition between you, on the table, is to prevent you influencing each other while you are scoring. I want your scoring to be independent. And please don't talk while you are observing the subject.

You may feel that I have told you so much that it will be difficult to remember, but it's quite simple. Score an X for a smile and an O for a nod. If there is an unbroken series of repeated nods just score one O. When the bell rings draw a line across the paper.

I'm sorry I can't offer to answer your questions but, if you are not sure what to do, use your best judgment.

Now, are you ready? Please do your best to score these subjects accurately.

Put your name on the end of the paper tape, . . . and, above your name, the number of the first subject. The number is "1", . . . "1" . . . What you have done, so far, should look something like this (Instructor held up a paper tape on which specimens of these items had been recorded in large writing). I hope you've got it right! . . . OK. Now, get ready to focus on the subject.

The first target subject was then shown on the screen for ten minutes. When he disappeared he was replaced by the instructor, who continued,

Well, I hope there were no difficulties. If you have not already done so, draw a line across the paper tape, beyond the last score, and tear the paper off beyond the line. Now, you can sit back and relax for 30 seconds, until you hear my voice again. (During the following interval of 30 seconds, as during all rest periods, the screen was blank).. . .

Now, let's get on with the next subject. Before we do, would you check that the front legs of your chair are at the marks on the floor? . . . Remember to draw the line across the paper when the bell rings at the half-way mark. Right, put your name on the end of the tape, . . . and, above your name, the number of the subject, which is "25", . . . "25". . . . OK. Prepare to focus on the subject.

As before, a target subject was shown for ten minutes and was then replaced by the instructor who continued,

> Right, draw a line beyond the last score and tear off the paper beyond the line. You should be getting good at the job! We'll have a one-minute break, so you can take it easy till you hear me speaking again. . . .
>
> Now, check the position of your chair. . . . Put your name on the tape, . . . and the number of the subject, which is "44", . . . "44" . . . This will be the last subject today. OK. Prepare to focus on the subject.

Once again a target subject was shown for ten minutes and was then replaced by the instructor who continued,

> Now, draw a line beyond the last score and tear off the paper beyond the line. . . . OK. You are finished for the day. Thank you very much for the work you have just done. Leave the paper tapes, that have the scores, on the table. Our next session will be a week from today, beginning at the same time, and you should come to the same waiting room. Before you go I would like to make a suggestion that could help to safeguard all of us. Would you mind exchanging phone numbers and phoning each other, as a reminder, the evening before the next session? You can tear off a piece of paper tape and make a note of the other's phone number now. . . . So, that's all for this time. I'll be looking forward to seeing you next week and, at the end of that session, your card will be signed for your two hours' credit.

This completed the instructions for the first session.

Instructions for observers scoring in AA Condition of Pairing differed from the above, in the first session, in that there was no mention of a partition, or of talking while the subject was on the screen, and there was no suggestion that phone numbers be exchanged. Observers scoring in PA Condition of Pairing received the same instructions as those in PP Condition, but at the end of the session the investigator arranged for the members of each pair to return at different times for the following session, the second observer being scheduled to come half an hour after the first.

Second Observational Session

The general procedure was similar to that employed in the first

session. Each observer remained in the Pairing condition to which he had already been allocated but Expectancy manipulations varied over observers. The following instructions were presented to observers in L Condition of Expectancy.

> I'm very glad you were able to keep your appointment. It's a great help, in an experiment like this, when everybody shows up.
>
> First of all, I'm going to remind you, very briefly, of what you have to do. Always keep focused on the subject, while he's on the screen, and don't look down at the paper you score on. When a subject smiles put an X. When he nods put an O. Score an unbroken series of nods as one nod. When the bell rings draw a line across the paper.
>
> Now, check the position of the front legs of your chair. . . . Put your name on the end of the tape, . . .and the number of the subject, which is "29", . . . "29" . . . We have finished now with the practice subjects. You have scored three of them, so you should be able toscore their smiles and nods quite accurately. There's one thing that will help you to score even more accurately. It has been found that accuracy is increased if observers know what emotional state the subject was in,—in this study if they know whether he liked or disliked the interviewer, or was neutral. So, I am going to tell you each subject's condition, and I want you to remember it while you are observing him. I also want you to write the subject's condition on the paper tape, beyond his number. Right, now remember that this next subject was in the Liked Condition. He liked the interviewer. You have already got his number. Beyond the "29" put the word "Liked." . . . What you have written will look like this (Instructor held up a paper tape on which specimens of these items had been recorded in large writing). OK. Prepare to focus on the subject.

Immediately after the first subject had been scored the instructor continued,

> OK. Draw a line across the paper beyond the last score, and tear it off beyond the line. . . . You can rest for 30 seconds before we score the final subject. . . .
>
> Now, are you ready for the last one? Check the position of the front legs of your chair. . . . Put your name on the end of the tape, . . . and the number of the subject, which is "2", . . . "2" . . . This subject was also in the Liked Condition. He liked the interviewer. So, write the word "liked" on the paper tape. . . . And try to remember, while

you are scoring, that this subject liked the interviewer. OK. Prepare to focus on the subject.

Immediately after the second subject had been scored the instructor continued,

> Now, draw a line across the paper tape, beyond the last score, and tear it off beyond the line. . . . Well, as I told you, that was the last subject. I hope you are not too tired, and that it has not bothered your eyes. Thank you very much for the way you have cooperated. It has been a great help to me.
>
> Don't forget to have your credit card signed before you leave. And, one last thing! The experiment will extend over two terms, and it will be about six months before the results are available. I think most of the procedures in the experiment are fairly obvious but, if you wish, I will be very glad to send you a summary of the experiment and its outcome. If you are interested, and will write your name and address on the envelope that my assistant will supply, I will mail the information to you. But, as I said, it will take some time.
>
> My assistant may want to speak to you for a minute but, apart from that, you are through. Good bye, now. And, again, thanks very much.

Instructions read to observers in N and D Conditions of Expectancy differed from those for L Condition only in having "Neutral" and "Disliked," respectively, substituted for "Liked," with the necessary related modifications.

Study II

First Observational Session

In this session all observers received the same instructions as were presented to observers in PP Condition in the first session of Study I. The investigator arranged for observers in PA Condition, and the confederate, to return at different times for the second session, as he did for the observers in this condition in Study I. The confederate, of course, did not in fact return.

Second Observational Session

In this session all observers received the instructions presented to observers in D Condition in the second session of Study I.

APPENDIX B

POST-EXPERIMENT QUESTIONNAIRE

Your frank responses to the following questions will assist the experimenter in evaluating the results and in planning future research. For those items which do not require a written answer please put a tick on the appropriate line. Do feel free to add any additional comments which you think will be helpful.

1. The location of the observer, in relation to the TV screen,

   a) Was satisfactory _____.

   b) Could be improved by

3. The clarity of the TV picture was generally

   Satisfactory _____. Unsatisfactory _____.

3. Did the instructions

   a) Leave you uncertain about some aspect of your task?

   Yes _____. No _____.

   If "Yes", please explain.

   b) Give you a clear idea of the purpose of the experiment?

   Yes _____. No _____.

   State what you understand the purpose of the experiment to be.

(Over)

4. Did the task make unreasonable demands upon you? E.g., eyestrain through having to observe for too long at a time.

Yes _____. No _____.

If "Yes", please elaborate.

5. Do you think you could predict the results of the experiment? As a guess, would you expect that subjects who <u>liked</u> the interviewer would

a) SMILE more than _____, same as _____, less than _____

b) NOD more than _____, same as _____, less than _____

subjects who <u>disliked</u> the interviewer?

6. From your incidental observation did the frequency of any other nonverbal behaviors, on the part of the last two subjects strike you as surprisingly high or low?

The subjects'

_____ (behavior) were surprisingly high _____/low _____.

_____ (behavior) were surprisingly high _____/low _____.

APPENDIX C

DE-BRIEFING LETTER TO OBSERVERS

The Department of Psychology,
University of Manitoba,
Winnipeg 19, Manitoba.

July, 1971.

Dear

You may remember that, within the last twelve months, you took part in an experiment called either "Napuhsi" or "Mufhup" and that the experimenter promised to provide some information about the study and its outcome.

The investigation was concerned with influences which may affect the accuracy of observers used to score the responses of subjects in psychological experiments. One problem studied was the influence on observers' scoring of information or beliefs about the subjects. Groups of observers were given varying information. Some were told that the subjects liked an interviewer, some that they disliked him, and others that they were neutral, i.e., neither liked nor disliked him. All observers saw the identical subjects. It was found that observers who scored alone were influenced by this information. Those who had been told that the subjects liked the interviewer scored more smiles than either of the other groups.

The other problem concerned the possible influence of paired observers on each other's scoring. It has been the custom of investigators using observers to attempt to test their performance by correlating the scores of an observer with those of his partner, a high correlation, representing high scoring agreement, being taken as an indication of satisfactory scoring. Obviously if observers influence each other's scoring this could account, at least in part, for high agreement between them, as well as being a source of inaccuracy in their scoring. To examine these possibilities the observers were paired under three conditions. Some had a scoring partner in both practice and test sessions, some had a partner in the practice session but scored alone in the test session, while others scored alone in both sessions. The scores of paired observers were correlated. Correlations were also computed for the scores of the same observers randomly matched and also for random matchings of observers who scored alone in both sessions. It was assumed that if scores of paired observers correlated at a level significantly higher than the scores of the same observers randomly matched an inter-observer influence, or scoring consensus, was indicated. Such consensus was found for smiles when observers were paired in both sessions, and for nods when observers were paired in the practice session and scored alone in the test session. Correlations for randomly matched observers who

scored alone in both sessions were close to zero for both smiles and nods. The results indicated that scoring consensus did occur, in one or other of the paired observer conditions, for smiles and nods. Further analysis indicated that consensus occurred only when observers were told that the subjects disliked the interviewer.

As a further test for the consensus effect an additional experiment was conducted. Observers were paired with confederates, posing as observers, who scored consistently at a predetermined high or low level. All observers were told that the subjects disliked the interviewer. Observers paired with high-scoring confederates scored significantly more smiles and nods than observers paired with low-scoring confederates. This occurred for smiles when observers were paired in both sessions. The effect was found for nods irrespective of whether the observers were paired in both sessions or only in the practice session but was stronger in the latter condition. Thus, the consensus findings of the first experiment were closely replicated in the second and it was demonstrated that the consensus effect can be manipulated experimentally.

The results indicate that observers scoring alone may be influenced in their output by their knowledge or beliefs regarding the subjects they score. It is also evident that paired observers may have a considerable influence upon each other's scoring, and that correlating the scores of paired observers is an unsatisfactory test of their scoring accuracy unless very stringent precautions are taken to prevent consensus occurring.

You will now have realized that you were not told the true purpose or nature of the investigation, that the information you were given about the subjects' attitudes towards the interviewer conflicted with the information given to other observers viewing the same subjects, and that if you had a partner he may have been a confederate and not an observer. Unfortunately, this research could not be conducted without these unusual procedures. On the other hand, the outcome should provide valuable guidance for future psychological research and enhance the potential of psychology for humanitarian purposes.

I extend my apologies for not having been able to debrief you sooner on the procedures and results of this investigation, but the nature of the study rendered it inexpedient to do so. I would also like to repeat my expression of thanks for your cooperation in these experiments and to wish you every success in your own studies.

Sincerely,

Douglas Montgomery

APPENDIX D

Table 14

Observers' Awareness of the Purpose of the Experiment

Variable	L			N			D			Totals		
	U[a]	?[b]	A[c]	U	?	A	U	?	A	U	?	A
AA	16	0	0	16	0	0	16	0	0	48	0	0
PA	15	1	0	14	1	1	16	0	0	45	2	1
PP	15	0	1	16	0	0	16	0	0	47	0	1
Totals	46	1	1	46	1	1	48	0	0	140	2	2

[a]Unaware observers

[b]Observers whose responses were ambiguous and who may, or may not, have been aware.

[c]Aware observers

Table 15

Direction of Observers' Expectancies for Smiles

Variable	L			N			D			Totals		
	+[a]	0[b]	-[c]	+	0	-	+	0	-	+	0	-
AA	12	3	1	16	0	0	11	5	0	39	8	1
PA	14	1	1	10	6	0	12	3	1	36	10	2
PP	14	2	0	13	1	2	9	3	4	36	6	6
Totals	40	6	2	39	7	2	32	11	5	111	24	9

[a]Observers' expectancies were as anticipated.

[b]Observers either did not respond or indicated that the subjects' attitudes to the interviewer would not affect the output of smiles.

[c]Observers' expectancies were opposite to those anticipated.

Table 16

Analysis of Covariance of Observers' Expectancy and Pairing Effects for Smiles, First and Second Target Subjects

Source	df	First Subject		Second Subject	
		MS	F	MS	F
Expectancy (A)	2	13.251	0.51	3.859	0.18
Pairing (B)	2	1.473	0.06	12.553	0.59
A x B	4	28.849	1.10	43.635	2.04
Error	134	26.129		21.341	

Table 17

Adjusted Means for Observers' Expectancy and Pairing Effects for Smiles, First and Second Target Subjects

Variable	L	N	D	Total
		First Target Subject		
AA	14.314	12.672	11.558	12.853
PA	12.939	14.832	11.696	13.040
PP	12.545	12.028	13.517	12.690
Total	13.272	13.040	12.690	
		Second Target Subject		
AA	5.989	3.460	3.807	4.425
PA	5.177	6.967	4.161	5.437
PP	4.467	3.688	6.284	4.805
Total	5.214	4.694	4.759	

Table 18

Analyses of Covariance, for Smiles,

for Each Condition of Pairing

Source		AA		PA		PP	
	df	MS	F	MS	F	MS	F
Between	2	120.071	3.583*	140.122	3.696*	44.583	1.225
Within	44	33.509		37.912		36.386	

*$p < .05$.

Table 19

Analysis of Covariance for Smiles,

First and Second Periods

Source		First Period		Second Period	
	df	MS	F	MS	F
Expectancy (A)	2	18.611	1.34	3.507	0.13
Pairing (B)	2	7.664	0.55	1.883	0.07
A × B	4	18.421	1.32	71.753	2.66*
Error	134	13.925		27.012	

*$p < .05$.

Table 20

Adjusted Means for Observers' Expectancy and Pairing Effects, for Smiles, First and Second Periods

Variable	L	N	D	Total
		First Period		
AA	8.698	6.848	6.586	7.667
PA	8.788	8.723	6.880	8.125
PP	7.675	6.481	8.384	7.229
Total	8.391	7.341	7.287	
		Second Period		
AA	11.886	9.249	8.915	10.009
PA	9.414	12.700	8.789	10.299
PP	9.258	8.894	11.603	9.921
Total	10.181	10.275	9.774	

Table 21

Analysis of Variance of Observers' Expectancy and Pairing Effects for Smiles and Nods

Source		Smiles		Nods	
	df	MS	F	MS	F
Expectancy (A)	2	68.3958	0.96	295.1316	0.61
Pairing (B)	2	40.5833	0.57	1431.4644	2.96
A x B	4	203.0726	2.84*	789.4189	1.63
Error	135	71.4275		483.2117	

*$p < .05$.

Table 22

Means for Observers' Expectancy and Pairing Effects for Smiles and Nods

Variable	L	N	D	Total
		Smiles		
AA	21.31	15.69	16.50	17.83
PA	19.13	20.44	16.31	18.63
PP	16.00	13.25	21.13	16.79
Total	18.81	16.46	17.98	
		Nods		
AA	31.06	27.56	22.31	26.98
PA	14.31	31.56	21.94	22.60
PP	12.81	13.94	21.63	16.13
Total	19.40	24.35	21.96	

Table 23

Correlation Coefficients for Paired and Randomly Matched Observers, for Smiles, First and Second Target Subjects

Conditions of Pairing	First Subject		Second Subject	
	Paired	Random	Paired	Random
AA		.022		-.017
PA	-.047	.049	.263	-.014
PP	.329	-.029	.763[a]	-.135*

*$p < .0000005$ for difference between coefficients.

[a]$p < .01$ for correlation coefficient.

Table 24

Correlation Coefficients for Paired

and Randomly Matched Observers,

in Each Expectancy Condition, for Smiles

Expectancy	Paired	Randomized
Liked	.468[a]	.498[a]
Neutral	.502[a]	.533[a]
Disliked	.661[a]	.095*

*$p < .0001$, for difference between coefficients within row.

[a]$p < .005$, for correlation coefficient.

Table 25

Correlation Coefficients for Paired and Randomly Matched Observers for Nods, First and Second Target Subjects

Conditions of Pairing	First Subject		Second Subject	
	Paired	Random	Paired	Random
AA		-.233		-.053
PA	.359	-.156*	.442[a]	-.227**
PP	-.140	.069	.352	.022

*p < .01 for difference between coefficients.

**p < .001 for difference between coefficients.

[a]p < .05 for correlation coefficient.

Table 26

Correlation Coefficients for Paired and Randomly Matched Observers, in Each Expectancy Condition, for Nods

Expectancy	Paired	Randomized
Liked	.286	-.111*
Neutral	.231	-.175*
Disliked	.587[a]	.076**

*p < .02, for difference between coefficients within row.

**p < .001, for difference between coefficients within row.

[a]p < .005, for correlation coefficient.

Table 27

Analysis of Variance for Nods,

First and Second Target Subjects

Source	df	First Subject		Second Subject	
		MS	F	MS	F
Level (A)	1	1443.9961	20.34*	2575.5623	20.18*
Pairing (B)	1	18.0625	0.25	5.0625	0.04
A x B	1	68.0623	0.96	351.5625	2.75
Error	60	70.9936		127.6371	

*$p < .001$.

Table 28

Means for Nods, for First and

Second Target Subjects

Variable	Subject 1			Subject 2		
	High	Low	Totals	High	Low	Total
PA	15.62	4.06	9.84	26.81	9.44	18.13
PP	12.50	5.06	8.78	21.56	13.56	17.56
Total	14.06	5.56		24.19	11.50	

Table 29

Mean Confederate Effects

Variable	Smiles		Nods	
	Confederate		Confederate	
	A	B	A	B
HSC	17.44	16.94	43.88	32.62
LSC	10.91	13.94	13.75	18.38