JOHNSON, Vernon Scott, 1938-
THE BEHAVIOR THERAPY CONTROVERSY: A STUDY ON OBSERVER BIAS, RELIABILITY, AND INFORMATION.

Iowa State University, Ph.D., 1970
Psychology, clinical

University Microfilms, Inc., Ann Arbor, Michigan

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED
THE BEHAVIOR THERAPY CONTROVERSY: A STUDY ON OBSERVER

BIAS, RELIABILITY, AND INFORMATION

by

Vernon Scott Johnson

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Psychology

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Head of Major Department

Signature was redacted for privacy.

Dean of Graduate College

Iowa State University
Ames, Iowa
1970
# TABLE OF CONTENTS

## REVIEW OF LITERATURE
- Discussion
  - Learning theory
  - Applicability to psychotherapy
  - Experimental controls

## RATIONALE AND PROCEDURE
- Goals
- Hypotheses
- Procedure
  - Staff
  - Patient selection
  - Patient description
  - Resource people
  - Training of resource people
  - Collection of content data
  - Behavior change measures
  - Subjects
  - The rating scale

## RESULTS
- The Comparison of the Means
- The Reliability of the Items
- The Reliability of the Items for Rater Pairs
- Rater Preference Within or Between Treatment Groups

## DISCUSSION

## BIBLIOGRAPHY

## ACKNOWLEDGMENTS

## APPENDIX I: RATING SCALE

## APPENDIX II: INSTRUCTIONS

## APPENDIX III: TWO EXAMPLES OF RESPONSE SET BY RATERS
REVIEW OF LITERATURE

The 1967 edition of Bernard and Orlando's bibliography of behavior modification (1967) lists over 860 references to papers which the editors maintain "develop or illustrate specific methods of behavior modification which were conceived of within some learning theory framework" (p. 1). Thus, one might conclude that the validity of the behavior therapy approach would not be challenged due to its sheer magnitude. However, this has not been the case. Four articles have been published in the Psychological Bulletin since May, 1965, which have debated the efficacy of behavior therapy (Breger and McGaugh, 1965; Rachman and Eysenck, 1966; Breger and McGaugh, 1966; Wiest, 1967). A fifth article (Katahn and Koplin, 1968) examines the controversy as a paradigm clash.

An attempt will be made to reiterate these arguments and counterarguments and show how they have or have not been resolved. The controversy centers around three major questions. They are: 1) Are the "laws of learning" which the behavior therapists claim to use actually accepted by contemporary learning theorists? 2) Are the learning models and techniques as applied in treatment settings appropriate for understanding and treating psychosis and neurosis? 3) Are the studies reported by the behavior therapists sufficiently controlled to offer support for the extremely high improvement rates the behavior therapists report to obtain? The following discussion will show that only the final issue remains unresolved. The present study deals solely with the likelihood that observers do not bias their ratings.

The controversy began with the 1965 article by Breger and McGaugh.
They attack the behavioral modification model as being unscientific because the principles upon which it is founded are being questioned in the learning laboratories. They also question the efficacy of the model as being appropriate to treatment of neurosis and maintain that: 1) what the behavior therapists call behavior modification is not different from traditional therapy; 2) when behavior therapists employ their techniques they are doing so in such a general and undefinable manner that there is little resemblance between their behavior and the behavior of the learning technician in the laboratory; and 3) the behavior therapists have not produced sufficiently controlled studies to claim success. They suggest the behavior modification paradigm should be supplemented or replaced by concepts utilizing plans, strategies, or other conceptualizations from cognitive approaches to learning.

Rachman and Eysenck deny that a cognitive approach to therapy is feasible and suggest that

... when Breger and McGaugh have some actual applications to report, or have at least succeeded in showing how the major facts of neurotic behavior can be accounted for in terms of their scheme, then may be the appropriate time to take issue with the 'reformulation' (1966, p. 165).

In the second portion of their paper they: 1) state that Breger and McGaugh are not familiar with the literature and support their claim by listing specific references to articles that Breger and McGaugh have excluded from their review; 2) answer the charge that learning laws are nonexistent by reverting to "the data closest to reality" and take a nontheoretical stand relying entirely upon demonstrable phenomenon; and 3) support the behavior therapists' claim that they employ learning principles by listing references to specific case histories. The issue of experimental
control is dealt with by objecting to the sample of studies Breger and
McGaugh cite and listing references they had not reported.

In their reply to Rachman and Eysenck, Breger and McGaugh (1966) reiterate the three questions asked in the original article, believing that Rachman and Eysenck did not satisfactorily answer any of their criticisms. They accuse Rachman and Eysenck of ignoring the issues they raised about the deficiencies in S-R psychology. They point out that many of the techniques which were supposed to have been developed from S-R learning theory were, as a matter of fact, in existence prior to the development of S-R psychology. The references given in the Rachman and Eysenck rejoinder to the criticism of poor control are reviewed. Breger and McGaugh conclude "that our original criticisms are still fully applicable" (1966, p. 171).

On the issue of the learning model they again attack the S-R learning model and propose that its deficiencies can be overcome only by the addition of some form of cognitive processes. They specifically argue against the generality of learning and the appropriateness of the law of effect. They further maintain that peripheral or S-R psychology cannot explain the equivalence of responses and the adaptive application of behavior in novel situations without turning to a mediational concept or cognitive notions such as strategy, hypothesis testing, or plans. It is pointed out that the results of the perceptual consistency, stimulus equivalence, response equivalence, and transposition controversies are evidence that the effective stimulus is clearly more "central" than "receptor". The response equivalence and place versus response learning data are interpreted as suggesting that the organism is learning what needs to be done in order to achieve some final event rather than learning a specific sequence of mechanical
responses consisting of a stereotyped pattern of muscular contractions.
Summaries of the latent learning, perceptual consistency, imitation, lan-
guage learning and imprinting studies are quoted as evidence against the
law of effect.

The Hullian notion that drive reduction is necessary for learning to
occur is countered by reviewing the latent learning, sensory precondi-
tioning, novelty and curiosity motivation, and direct brain stimulation studies.

As the attack upon the drive reductionist is not applicable to indi-
viduals employing operant techniques, Breger and McGaugh discuss the circu-
larit y in the definition of reinforcement. They maintain that such defini-
tions make it impossible to determine if results are dependent upon the
client's awareness rather than something inherent in the reward. Breger and
McGaugh suggest that the latent learning and discrimination reversal studies
refute any suggestion that behavior can be understood when one knows the
reinforcement history of the organism as they show examples where perform-
ance is not dependent upon reinforcement.

Rachman and Eysenck defend their position by maintaining that the con-
licts Breger and McGaugh wrote about were theoretical conflicts and irrele-
vant to the methodology of behavior modification. Such phenomena as acquisi-
tion, generalization and extinction are, they maintain, laws of learning.
It is their contention that the controversies cited are conflicts over
theory, that no learning theorist would deny statements of behavioral laws
concerning the effects of reinforcement, extinction, etc., and that laws of
this type are used by behavior therapists. They suggest that the cognitive
versus S-R issues, i.e., perceptual consistency and the concept of rein-
forcement, are adequately summarized by Taylor and Papert (1956), Taylor
(1962), Mowrer (1960), and Metzner (1961; 1964), and need not be reconsidered in their article.

The second Breger and McGaugh article (1966) asserts that Rachman and Eysenck have not answered the criticisms leveled at the adequacy of the S-R model in explaining perceptual consistency, language interpretation and generation, and response equipotentiality. They claim Rachman and Eysenck have avoided the issue by retreating to "findings" from conditioning studies, i.e., acquisition, extinction, etc., which are entirely dependent upon the conditions under which they were studied and should not be generalized to a nonlaboratory setting. They further suggest that there is little similarity between the behavior therapist and the learning technician in actual context and that it is inconsistent for behavior therapists to make claims of widespread behavioral changes as a result of their treatment while they limit their theory to highly specific laboratory conditions.

In 1967, Wiest published a paper in which he accuses Breger and McGaugh of misrepresenting the S-R position and then attacking the obvious fallacies inherent in their misrepresentation. His purpose is to clarify the misinterpretation and to show that, if interpreted correctly, S-R psychology is not faced with many of the deficiencies represented by Breger and McGaugh's criticism.

He begins with their definition of how a behaviorist defines learning, i.e., "The tendency to make a particular response in the presence of a particular stimulus". He claims the statement is absurd when one considers any unconditioned reflex, or the instinctive behavior that the ethologist describes as released by sign stimuli. He maintains that the statement is not representative of a behavioristic standpoint, but that it does go to
"show how easily a straw man can be felled" (1967, p. 215).

Since Breger and McGaugh had denounced the S-R psychologist for not dealing with many of the controversies in the learning laboratory, Wiest spends some time explaining that:

... it is not true ... that any set of observations (e.g., the observations referred to by the terms 'transposition', 'latent learning', 'response equipotentiality', etc.) implies a particular theory .... Those phenomenon that cognitive theorists believe are best explained by a theory with cognitive intervening variables or mediating processes can also be conceptualized in terms of quite different (noncognitive) theories. It would seem advisable, therefore, not to form constructs about some phenomenon in terms of our own theory and then criticize other scientists for not dealing with our constructs. The only things that the other scientist must deal with are publicly observable facts ... it is not incumbent upon him to deal with our inferences (1967, pp. 219-220).

Wiest points out that in their discussion of reinforcement, Breger and McGaugh's choice of the saying, "one man's meat is another man's poison" was not valid since "... if the saying were true, man as a social animal, with his potlatches and potlucks would have descended from a long line of extinct ancestors" (1967, p. 216). On a more serious note he addresses the issue that Breger and McGaugh claim the reinforcement theorist professes to be able to predict behavior from a knowledge of the reinforcement history. Wiest notes that both reinforcement and cognitive theorists make use of an organism's reinforcement history in making any predictions, but neither claims there are no other determinants of behavior.

On the issue of discrimination reversal experiments Wiest explains that the behavior in question does, in fact, conform to the reinforced contingencies: the organism does change his response when the formerly correct response fails to pay off.

In summary, Wiest's main lines of argument are: 1) Breger and McGaugh
are misinformed; 2) cognitive theorists frequently are working within a different paradigm than S-R theorists, i.e., one group should not expect their opponents to address alien theoretical constructs; 3) Breger and McGaugh have tended to label a behavioral phenomenon and then proceed as if the label were an explanation of the phenomenon; 4) a large part of the confusion has been created by confusing response and stimulus and then attributing the confusion to the behavior therapist or the S-R psychologist.

Discussion

The purpose of this review is not to debate the issues that have been presented but rather to examine the questions raised as a prelude to a study that will attempt to evaluate one of the issues.

If one evaluates the arguments and counterarguments that have been raised, it is apparent that in the main part the issue is a carryover from the unresolved, and perhaps unresolvable, issues which have prompted the mechanistic versus humanistic debates for the past 2000 years. In the following pages the status of the S-R learning theory, the applicability of the learning model to psychotherapy, and the lack of controlled studies will be discussed. An attempt will be made to reach a conclusion about each.

Learning theory

The issue, as outlined by Breger and McGaugh's section on learning theory, is not that the techniques of the behavior therapist are inappropriate but rather that their claim to using techniques derived from "learning laws" is unjustified. Their criticisms are: 1) techniques similar to behavior modification predated the advent of S-R psychology, i.e., the development of behavior therapy is independent of the events which occur in
the learning laboratory; and 2) the laws of S-R psychology are debatable as laws of learning.

Breger and McGaugh relate that:

Pfaundler (1904) and Nye (1930) both describe methods highly similar to the behavior therapist techniques that were employed before the advent of S-R psychology and for that matter, circus trainers have been using 'operant techniques' for centuries (1966, p. 172).

This is the basis for the argument that what goes on in learning theory has contributed little to the development of behavior therapy techniques.

It would be unreasonable to argue against the obvious conclusion that what are called behavior modification practices were present prior to the advent of psychology, much less the occurrence of S-R learning theory. However, one would be well advised to question why the procedures failed to find substantial acceptance within the practice of the clinician until the area of learning theory had been established. It is begging the issue to deny that the conceptual framework from which behavior modification operates developed in the learning laboratories simply because there are examples where similar techniques were employed before S-R psychology developed. If a behavior therapist wishes to maintain that he has developed an idea from a learning theory, disproof becomes impossible.

Weist states that the problems of transposition, latent learning, response equipotentiality, etc., are both statements of a particular theory and descriptions of empirical findings. He maintains that the scientist is required to explain only the empirical findings and not the theoretical construct.

His second general line of criticism pertains to the idea that the valid portions of the Breger and McGaugh attack were limited to concepts
which S-R theorists questioned years ago, i.e., the attack upon drive reduc­tion, Hull's r_g, and the idea that an animal in Pavlovian conditioning apparatus is doing more than salivating. It is his belief that if one looks at more recent S-R theory, he will find that the criticisms Breger and McGaugh list are no longer applicable.

In addressing the equipotentiality issue, Wiest reminds the reader that, "Some time ago, Skinner (1935; 1938) pointed to the generic nature of both stimulus and responses, and defined an operant as a class of responses" (1967, p. 216).

Rachman and Eysenck's claim that the concepts of acquisition, extinction, etc., are accepted by psychologists, independent of their theoretical orientation is justified. It is the status of the behaviorists' claim to laws of learning, not their techniques, which is challenged. Breger and McGaugh state this clearly when they say:

The reference to 'laws of learning' that 'no learning theorist of any persuasion would deny' further avoids the issue. What Rachman and Eysenck cite are findings from conditioning studies which are highly dependent on the conditions under which the observations are made. Although it may be that no learning theorist would deny that 'reinforced pairings of CS and UCS under appropriate conditions produce conditioning', many would deny that this finding is a 'law of learning' ... (1966, p. 170).

However, the accusation that behavior therapists are resorting to experimental findings and not learning laws will encounter little disagreement. Yet it is ironical that a student of learning, regardless of theoretical bent, would be willing to chastise a colleague for not having "laws" which are agreed upon by all contenders. If Breger and McGaugh's point is that the discipline of learning is not ready to deliver the final and undeniable statement whereby the study of learning is completed, then total agreement
from most psychologists is imminent. If, however, their point is that
learning psychologists do not have sufficient knowledge to make predictions
of the outcome of conditioning studies, their argument will not receive
general approval. The question, Are the behavior therapists' procedures
based on laws of learning or experimental findings? is academic. The rele-
vant issue is: would the shift of their status from experimental findings
to laws change the operations of the therapist? The answer is obvious.

Applicability to psychotherapy

The argument against the applicability of the behavior modification
model to psychotherapy has two main points. First, the behavior therapists
are not following the rigorous procedures of the learning theorist by de-
fining terms. Second, the operations of the behavior therapist are fre-
quently no different from more traditional therapists nor are they consist-
et with the behaviorists' stated position.

Breger and McGaugh (1965) maintain that in order to make the S-R model
and techniques applicable to neurosis, it is necessary to define the terms
in such a general way that they lose their meaning. They write:

When we look at the way conditioning principles are applied in the
explanation of more complex phenomena, we see that only a rather
flimsy analogy bridges the gap between such laboratory defined
terms as stimulus, response and reinforcement and their referents
in the case of complex behavior. Thus, while a stimulus may be
defined as an electric shock or a light of a certain intensity in
a classical conditioning experiment, Bandura (1951) speaks of the
'imagination of a scene', or, while a response may consist of
salivation or a barpress in a conditioning experiment, behavior
therapists speak of anxiety as a response (1965, p. 333).

Rachman and Eysenck fail to answer the criticism that the definitions
are too vague to be admissible to a learning laboratory and revert to using
quotations from Breger and McGaugh's article showing vagueness. They do
not defend the position of the desensitization therapist that imagining the items on a hierarchy constitutes a stimulus.

Wiest answers the question of how the desensitization therapist considers imagination a stimulus by pointing out:

The statement is misleading because it confuses stimulus and response and attributes the confusion to S-R theorists. Imagination is a form of behavior, not a stimulus (cf. Skinner, 1963) (1967, p. 216).

Hence, the confusion created by Breger and McGaugh is simply another example of their creating "straw men".

Despite Wiest's defense against the quotation used by Breger and McGaugh, there are examples of poorly defined terms in the behavior modification literature, e.g., "temper tantrum" (Williams, 1965), "eating behavior" (Bachrach, Erwin, and Mohr, 1965). The lack of clarity of a definition would be related to the reliability of the judgments concerning it and not to its adequacy as a construct. More will be said about this in a later section.

Breger and McGaugh complain that behavior therapists act in a way which is inconsistent with their position in that they interact with the clients much as more traditional therapists do and frequently use "dynamically" oriented descriptions of clients and outcomes. Their argument is supported by quoting from Rachman's writings. Rachman and Eysenck's answer to this is to agree that there is similarity in that a behavior therapist talks to the client and attempts to find out what events provoke or maintain the behavior and in other ways may do things similar to other therapists. However, they deny that the use of dynamically oriented descriptions are representative of behavior therapists and show that the statements Breger and
McGaugh quote in their article are taken out of context and are not representative of behavior therapy.

Wiest adds the criticism that the Breger and McGaugh paper is so vague in many of the criticisms about behavior therapists that it fails to make a distinction between behavior modifiers and more traditional therapists. He maintains the suggestions that behavior therapists account for present behavior in terms of generalization and that behavior therapists must posit a specific learning situation for each symptom could be criticism of a therapist of any theoretical orientation.

Both the experimenter and subject behaviors exhibited in behavior therapy are different from barpress training in complexity. However, the processes of extinction, acquisition, etc., remain. That the complexity of the response may vary as one encounters subjects of greater response capacity should surprise no one. It appears ludicrous to criticize the behavior therapists for applying instructions that save time. After all, their aim is to modify maladaptive behavior, not demonstrate that they can use conditioning to produce the same results as instructions but in a much less efficient manner. Whether the subjects are employing strategies, labels, generative processes or simply performing a conditioned response has no bearing upon data concerning frequency of response. The operations which result in changes in rate of performance do not themselves change as a result of the experimenters obtaining answers to the questions asked by the cognitive theorist.

Criticizing an experimenter for employing different forms of reinforcement or different levels of instructions for human than nonhuman subjects is not justified. If one examines the Premack principle, a consistency
among reinforcers becomes apparent even though the forms of reinforcers vary greatly for individual subjects (Premack, 1965).

**Experimental controls**

Much of the bitter ink that has gone into the controversy over the control issue could have been saved by recognizing that Breger and McGaugh are arguing about a theory of neurosis while Rachman and Eysenck are making references to techniques. Breger and McGaugh maintain that the claims of the behavior therapists are made on poorly controlled studies that are highly subject to sampling bias, observer bias, and lack of control. After reviewing the behavior modification literature cited in the reviews of Bandura (1961) and Grossberg (1964), it is their contention that only the Lang and Lazovik study (1963) demonstrates sufficient control to be considered theoretically significant and it deals with normal subjects from which results cannot be generalized to a neurotic population. They charge that behavior therapists typically select only cases which fit their theory and overlook "neurotic depressions, general unhappiness, obsessional disorders, and the kinds of persistent interpersonal entanglements that characterize so many neurotics" (1965, p. 348). They maintain that this is sampling bias in that it eliminates cases requiring a mediation explanation for understanding.

Observer bias, as used by Breger and McGaugh, occurs when judges are not "protected" from their own biases. From their analysis of 26 studies in their review, 12 are evaluated by persons other than the experimenter. Of these 12, four used hospital staff, four were parents, one a wife, and only three used an observer other than the therapist. It is their contention that "whatever factors enter in to cause observer and reporter biases
are allowed full rein in most cases" (1965, p. 352).

Breger and McGaugh are lenient in their allowances for experimental control in psychotherapy experiments, saying,

Since psychotherapy is not a controlled experiment, it is probably unfair to expect this type of control. However, there are more and less accurate descriptions of what goes on during any form of therapy, and we can demand as accurate a description as possible in lieu of experimental control (1965, p. 353).

Their criticism is that the behavior therapists have made no attempt to restrict the procedures in behavior therapy to learning techniques.

As previously mentioned, the argument is not against the techniques of the behavior therapists but rather against their claim to techniques derived from learning models. The acceptance of the technique is apparent in the following quotation:

While most of these techniques have been superceded by the various forms of dynamic psychotherapy, recent work (Frank, 1961) suggests that the time may be ripe for taking a fresh look at a variety of methods such as hypnosis, suggestion, relaxation, and other approaches of a more structured nature in which the therapist takes a more active role. Needless to say, this fresh look would best proceed unencumbered by an inadequate learning theory and with some minimal concern for control. ... there seems to be some evidence that these techniques (as techniques and not as learning theory) are effective with certain conditions (1965, p. 354).

Rachman and Eysenck, who operate from the behaviorist "no theory" position, answer these attacks against what Breger and McGaugh see as the behavior therapists theoretical position as though they were against their methods (1966, p. 168). They reiterate that the 1965 review of the studies done by Breger and McGaugh is based on a survey of all the studies reported in two reviews that have appeared (Bandura, 1961; Grossberg, 1964). It is their argument that behavior modification began around 1958-59 and that even a 1964 survey of the literature would be deficient as there had not been
time to develop and publish research. They then proceed to list references which Breger and McGaugh had not reviewed. They conclude:

The currently available evidence will, we feel certain, convince all but the most biased workers that the methods of behavior therapy are indeed effective in the modification of neurotic behavior (1966, p. 167).

Rachman and Eysenck (1966) answer the criticism that behavior therapists treat only highly selected cases with specific disorders by referring to articles specifying cases which were either selected by someone other than the therapist or were suffering from the specific maladies mentioned by Breger and McGaugh's criticism. They cite seven articles and a journal, founded since 1964, which they believe would have made the criticism of observer bias untenable. To add to the list they cite five references published prior to the Breger and McGaugh article which they claim show adequate experimental controls.

After reviewing the articles suggested by Rachman and Eysenck, Breger and McGaugh (1966) maintain that in the Eysenck volume (1964) which was listed as a resource for controlled studies, only two of the 42 articles are sufficiently controlled. These are the Lang and Lazovik article previously mentioned and an article by Anker and Walsh (1961) which makes no mention of learning theory but rather gives a good example of an innovative approach to patient management. The remaining articles listed in the Eysenck volume are claimed to show some form of experimenter, observer, or sampling bias. The conclusion drawn is that the:

... new references are as subject to bias as those previously cited, and the theoretical treatment of issues in the field of learning remains naive and misleading (1966, p. 172).

In a recent review of the literature on desensitization, Paul (1969)
reports the results of 75 papers which employed 90 therapists and over 1000 clients. Of the 75 studies he found:

... ten reports involved designs which could potentially evaluate specific cause-effect relationships between systematic desensitization therapy and change in client's behavior, although two of these ... fail to provide information because of methodological problems. Of the latter reports, the work of Lang and his colleagues ..., Paul ..., Moore ..., and Davison ... were all sufficiently controlled experimental studies to rule out within-class confounding, and all found solid evidence for the specific effectiveness of systematic desensitization (1969, p. 145).

Unfortunately, Paul (1966), Lang et al. (1966), and Davison (1965) were working with college students who could be considered normal. Only the paper by Moore (1965) deals with clinic patients, and these were asthmatic patients. Thus, the behavior therapists are still not in a position to point to their scientific achievements with neurotic or psychotic clients.

Three lines of questioning have prevailed throughout the controversy examined in this chapter. They are: 1) Are the claims of the behavioral psychologist to laws of learning justified? 2) Are the learning model and techniques applicable to neurotic and psychotic behavior? and 3) Are the studies offered as evidence of the success of behavior modification sufficiently controlled to warrant acceptance as either theoretically valid constructs or worthwhile methods?

It has been shown in the preceding discussion that there are defenses from the S-R camp for the theoretical issues raised and that a decision differentiating the learning model as based on laws or findings would not affect the operations or methods of the behavior therapist. In summary, the issue concerning the "laws of learning" as divergent from experimental findings is, for the behaviorist, irrelevant at best and meaningless at worst.

The arguments raised about the applicability of the behavioral model
to psychotherapy have been answered by Wiest as well as Rachman and Eysenck.

What has not been made clear, either here or in the preceding literature, is that the issue of the applicability of the behavioral techniques to treatment must rest on empirical demonstration of the validity of the technique. The exchanges have had little meaning if separated from the issue of experimental control. If, and only if, the behavior therapists' research is valid can the issue be resolved.

Much of the confusion has resulted from the failure of the opponents to see that one is talking about theory while the other speaks of methods. This leads to an unresolvable stalemate until the debate over the experimental control issue is resolved.

The appropriateness of the research reported by the behavior modifiers was shown to hinge on the issue of observer bias. There were three general criticisms of the literature: 1) The control procedures in an applied setting are not rigid. Breger and McGaugh capitulate on this issue as the type of control they speak of is simply not feasible with human subjects in a therapeutic setting. 2) The behavior modification studies are not controlling for sampling bias. 3) The behavior modification studies are open to criticism of observer bias. The criticisms are justified. However, in view of the number of controlled studies which suffer from the possibility of observer bias, but not sampling bias, the relevant issue becomes one of observer bias. The present study was carried out to investigate this issue.
RATIONALE AND PROCEDURE

The charge that the behavior modification literature is plagued with the possibility that problems of experimenter, observer, or sampling bias may have occurred are well founded. Despite the previously cited work of Paul (1966), Lang and Lazovik (1963), and Davison (1965), there is doubt that any single controlled study shows the effects of behavior change on a nonnormal population independent of sampling bias, or rater/observer bias. However, the failure to safeguard against the possibility of bias and the actual occurrence of bias are separate issues. A failure to safeguard against bias does not necessitate the occurrence but simply suggests that one cannot definitely say whether or not bias did occur.

If one examines the three forms of bias more closely, it will become apparent that the relevant issue is that of observer bias. On the question of experimenter control Breger and McGaugh simply make a case for description of methods rather than control. There are studies in the literature which are not subject to sampling bias. However, as Breger and McGaugh have shown, these studies have not controlled for observer bias. It would seem that the issue of sampling bias would be resolved if one can indicate the low feasibility of observer bias having affected these experiments.

If the criterion used in behavior therapy were shown not to be susceptible to biasing by the experimenter, hospital staff, etc., the statement, "Whatever factors enter in to cause observer and reporter biases are allowed full rein in most cases", and the criticism of these studies based upon that statement become important (Breger and McGaugh, 1965, p. 352).

Behavior therapists predominantly have employed measures which they
consider objective. The data are usually reported in terms of increases or decreases in the frequency of a response or in shifts on rating scales. If one could make the assumption that the count or rating was valid, the charge of observer bias loses its effectiveness. If, however, the measures are susceptible to bias, the criticism stands to gain strength.

Goals

The present study attempted to: 1) determine if observer bias will be obtained in a behavioral setting where ratings are done on mental patients, 2) construct a scale to estimate the approximate range of reliability required to diminish the influence of rater bias, 3) study an aspect of the relationship between rater bias and knowledge of the experimental conditions, and 4) demonstrate a method for assessing observer bias.

Hypotheses

Five hypotheses were investigated. They were:

1. Bias will occur when staff members who have worked with experimental groups rate the members of those groups.
2. Bias will be limited to items having low reliability.
3. Bias will not occur when the rater is aware of the possibility of bias but unaware of which individuals are involved in each treatment group; these conditions will lead to a lack of reliability between raters.
4. Bias will be prevented and reliability of the ratings will be preserved when the raters know the reliability of the items employed as well as which subjects are experimental and control.
5. Bias will not be evident in the highly reliable items; these items will not be influenced by any of the conditions of the experiment.

In order to test these hypotheses it was necessary that raters be asked to evaluate or rate subjects who were engaged in a behavior modifica-
tion project. By providing the raters with behavior checklists composed of items of varied but known reliability and manipulating the amount of knowledge about the experiment that raters had available, it was possible to determine whether involvement with the research could influence the ratings in the manner suggested by Breger and McGaugh (1965; 1966). In order to accomplish this, it was necessary to establish a behavior modification setting in which ratings were an integral part of an evaluation system used on nonnormal subjects who had been involved in a behavior modification program.

Procedure

Ten individuals from the community resource agencies were asked to explain the relationship that a veteran might develop in their organizations. Each individual made two presentations, one utilizing a role playing or practice technique and the other utilizing a more traditional lecture paradigm. Two staff members were present during each of the presentations. In the practice group these staff members served to give feedback or verbal reinforcement to the subjects engaged in the role playing. In the lecture group they simply observed and on occasion asked questions of the lecturer. It was felt that the practice group offered: 1) an opportunity for reinforcing the patient as he role played the interviewer and 2) an opportunity for social modeling to occur as patients witnessed a member of their group interact with the resource person.

Data would be collected on two separate dimensions. The first data were to assess changes in the patient's knowledge about the organization represented by the resource person. That is, does the patient learn from
the presentation? Second, data were collected and analyzed for changes in the patients interviewing skills.

Ten sessions were scheduled for two groups of patients. The two groups met weekly for an hour. They were scheduled to start 30 minutes apart. The first 15 minutes of each hour was devoted to obtaining a pre-exposure content measure. During this time the staff visited informally with the patients about the topic for the day. At the end of the 15 minute period the resource agent was brought into the room and immediately began his presentation. After 30 minutes if the resource person had not completed his presentation, he was interrupted and left the room. If he had covered the material in less than the 30 minute period, the remainder of the time was used as a question and answer session. The last 15 minutes of the hour was reserved for the final content questionnaire followed by the staff answering questions. During sessions one through five the practice group had the first contact with the resource agent. The order was reversed for the final five sessions.

Staff

The staff directly involved in the training sessions consisted of a female occupational therapist, a male chaplain and two male psychology trainees. All had approximately seven months of experience in working with role playing groups and were practiced in using verbal reinforcement techniques.

For each training session the four staff members were divided between the practice and lecture groups. Two members of the staff would work with the resource agent on either the practice or the lecture presentation.
Those individuals had from nine months to seven years experience working with the patients and were engaged in the study as a portion of their regular duties. They were assigned to teams in a counterbalanced fashion so that each of the four staff members worked with each group an equal number of times.

**Patient selection**

Thirty-six patients were originally selected. The criteria for selection were approval of the hospital staff, including four persons not involved with the experiment, based upon their subjective feeling that: 1) the patients would not be disturbing to the meetings; 2) the clients selected for participation have a reasonable chance of remaining in the group for the duration of the experiment, i.e., that there be no plans in operation for sending the subject into the labor market or a job training setting; and 3) the clients' problems were primarily behavioral and not physical (brain damage). As the study was done in an open setting which the patients attend by choice, it was mandatory that the clients be willing to participate in the meetings. Of the 36 men suggested by the staff members a total of 30 were willing to engage in the activity and to take the necessary test for evaluation for the project. These individuals were listed in the order they were contacted by the staff. From the list they were assigned randomly to treatment groups of 15 subjects to each group. Of these, two dropped out before the completion of the project. One obtained employment, and one simply found the interaction too threatening and ceased to come to the meetings.
Patient description

A survey of the population from which the patients were drawn was conducted six months prior to the onset of this project. It showed the typical patient to be a veteran of World War II who served in either the Army or the Navy. He had limited job skills (53% unskilled or skilled labor), no wife or children, lived in a foster home, and had come from a moderate sized community. He had probably been in a mental hospital on more than one occasion for a total of seven to eight years of hospitalization. The chances were even that he had a guardian, i.e., was declared legally incompetent. Two diagnostic categories account for 80 percent of the population -- chronic undifferentiated schizophrenia and paranoid schizophrenia (Nord, 1970). It is assumed that the patients selected for the study were representative of this population.

Resource people

Ten individuals from a variety of professions in the immediate area were asked to participate in the project as resource people, or visiting lecturer-role playing participants. They were drawn from government agencies, educational institutions, industry and the ministry.

Training of resource people

Each of these individuals was given a brief training session in the role playing technique. This was typically a rehearsal of their role with the chaplain working on the project and usually consisted more of providing reassurance that the role playing would be little different than the interaction ordinarily expected of their profession.
Collection of content data

Immediately before and after each session with the resource agent the veterans filled out a multiple choice test composed of content items which would be covered in the resource agent's presentation. The content items for the tests were written by the participating staff members. The test usually required less than ten minutes for completion.

Behavior change measures

On three occasions during the project a simulated job interview was staged using the participants from the project: once prior to the assignment of subjects to groups, a second time midway through the project, six weeks after the first rating session, and the third time one week following the final session with a resource agent. The subject's behavior was rated on a zero to nine scale composed of 25 items. The content of the items ranged from counting data (The number of questions subjects asked the interviewer) to inferential statements about motivation and ego strength.

Subjects

During the rating sessions the staff working with the patient groups were joined by four other raters. This provided eight separate ratings for each patient interviewed. The eight raters were divided into four groups or pairs.

The first pair consisted of two paid psychological consultants whose contact with the project was limited to their ratings. They did not know: 1) which patients were assigned to the separate treatment groups, 2) the purpose of the experiment on comparison of methods, or 3) that their rating of patients would serve as comparisons for a study of rater bias. These
subjects will be referred to as double blind (BB).

The second pair, consisting of two Ph.D. counseling psychologists employed by the hospital but not directly connected with the patients involved in the study, knew: 1) that they were rating patients which had been trained in either a role playing or lecture group, and 2) that their ratings would serve as comparisons in a study of rater bias. They did not know which of the patients belonged to the practice or the lecture treatment groups. This pair will be called single blind (B).

The third pair consisted of two interns who were completing their doctorate degrees in psychology. These individuals were aware of all contingencies as they: 1) were involved with both treatment groups, 2) were aware that their ratings would be used in a study of experimenter bias, 3) knew the reliability of the items used, and 4) had designed both experiments. These raters were considered double wise (WW) as they knew the purpose of the present study and the patients in each treatment group.

The final pair of raters, the single wise raters (W), was composed of a chaplain and an occupational therapist who were under the impression that their ratings were to be used to evaluate the comparison of the practice and lecture groups' performances. They were further given the information that the practice group was showing a superior performance on the content measure and were led to believe that if the behavior ratings showed a significant difference in favor of the practice group, they would be virtually assured of publishing the results.
The rating scale

The rating scale consisting of 53 items was written for this study (see Appendix I). These items were used by five Drake University students from a class in Abnormal Personality to rate simulated job interviews. The students playing the roles were coached by one of the psychology interns to insure variance in performance between role players. They were then interviewed by the other intern. The raters were allowed five minutes to complete their rating scale before the next simulated interview was begun. The inter-rater correlation for each item across the individuals who engaged in the role playing was obtained. From these 25 items were selected for the final rating scale (see Appendix II).
RESULTS

The Comparison of the Means

Had rater bias occurred in the present study, it would be apparent in a comparison of the means for the four pairs of raters (BB, B, WW and W). These raters, unfamiliar with the patient assignment to treatment groups, do not have the information required to bias their ratings. Therefore, the blind and double blind raters' mean ratings serve as the basis for comparison with the mean ratings from the wise and double wise rater pairs. Inspection of Figures 1 and 2 clearly shows that the mean ratings for the wise and double wise raters lie within the range of the blind and double blind raters. This precludes the necessity of testing for significance for obvious reasons.

Further, inspection of Tables 1 and 2 also indicate it is doubtful that there was a systematic effect from knowing that the ratings could be analyzed for rater bias. No rater pair occupies the same rank position for the magnitude of their means on the second and third rating session. If knowledge of patient assignment to group were to introduce some form of bias, it would not be expected to reverse itself between the rating sessions.

Figures 1 and 2 clearly indicate that no rater bias occurred which could have affected conclusions based upon the analysis of the data comparing the treatment techniques. Other forms of bias could occur and will be discussed in the following pages.

The Reliability of the Items

It will be recalled that a 25 item scale was constructed which varied greatly in the reliability of the items. Individuals from the patient
Figure 1. Mean ratings for rater-pairs on the 2nd evaluation session, Practice and Lecture groups combined.
Figure 2. Mean ratings for rater-pairs on the 3rd evaluation session, Practice and Lecture groups combined
Table 1. Reliability estimates for the rating scale items (decimals have been omitted)

<table>
<thead>
<tr>
<th>Item #</th>
<th>Rating 1</th>
<th>Rating 2 &amp; 3</th>
<th>Item #</th>
<th>Rating 1</th>
<th>Rating 2 &amp; 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>58</td>
<td>13</td>
<td>41</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>64</td>
<td>43</td>
<td>14</td>
<td>47</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>27</td>
<td>15</td>
<td>23</td>
<td>--</td>
</tr>
<tr>
<td>4</td>
<td>47</td>
<td>54</td>
<td>16</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>09</td>
<td>-- b</td>
<td>17</td>
<td>29</td>
<td>--</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
<td>49</td>
<td>18</td>
<td>35</td>
<td>44</td>
</tr>
<tr>
<td>7</td>
<td>27</td>
<td>19</td>
<td>19</td>
<td>16</td>
<td>--</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>31</td>
<td>20</td>
<td>20</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>40</td>
<td>39</td>
<td>21</td>
<td>40</td>
<td>36</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>--</td>
<td>22</td>
<td>51</td>
<td>--</td>
</tr>
<tr>
<td>11</td>
<td>32</td>
<td>--</td>
<td>23</td>
<td>35</td>
<td>--</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>--</td>
<td>24</td>
<td>35</td>
<td>30</td>
</tr>
</tbody>
</table>

a The correlations for rating sessions 2 and 3 have been averaged together to provide a single reliability estimate.

b Items having zero variance have been omitted.

The population previously described were evaluated by means of this scale on three separate occasions as they engaged in a simulated job interview. The ratings were done by the eight raters described as subjects. They used a nine point Likert scaling procedure. The data were scored and punched on data cards by an IBM 1230 optical scanner scoring machine.
A correlation matrix was obtained between raters for each item across the patients rated on the first rating session. Each correlation is an inter-rater reliability, i.e., the correlation of two raters over 22 patients. The average of these correlations indicates the reliability of the items for the eight raters used as subjects in the present study (Nunnally, 1970). It will be recalled that the patients had not been assigned to treatment groups at the time this measure was taken. Therefore, it is assured that no observer bias due to treatment groups can be present. The reliability estimates are given in column one of Table 1.

The Reliability of the Items for Rater Pairs

Campbell and Fiske (1959) suggest that individuals make similar scores on a test as a result of either the individuals tested having shared a common characteristic (trait) or from something inherent in the process of testing (method). They have shown that by forming four possible combinations of correlations it is possible to determine whether the similarity among tested individuals is due to methods or traits. By forming groups of correlations which share: 1) common methods and traits, 2) common methods but different traits, 3) different methods but common traits, and 4) different methods and different traits they can isolate the subject similarity due either to the traits or the testing methods. The analysis used in this study is a modification of their procedure.

The data collected on the second rating session were correlated across the eight raters for the two treatment groups. This resulted in a 16 x 16 matrix for each item or a total of 25 matrices. Borrowing terminology from Campbell and Fiske, one could envision the matrices as composed of two
methods and eight traits in which the methods would be the two treatment groups and the traits the raters. The average correlation for the portion of a matrix corresponding to a monomethod-heterotrait triangle would be the reliability estimate for an item for the eight raters within a treatment group. The portion corresponding to the heteromethod-heterotrait block results from the arbitrary pairing of patients in the lecture group with patients in the practice group. This portion of the matrix is not relevant for the present study.

The data for the third rating session were correlated in the same fashion, again yielding 25 matrices each having 16 rows and 16 columns. Because 12 of the matrices generated for the rating sessions contained an item which was assigned zero variance by one or more of the raters, there were no correlations for that rater in that group. These matrices were dropped from further analysis.

Three sets of operations were done on the remaining matrices: first, the correlation coefficients in the monomethod-heterotrait triangles were averaged for each pair of raters, rounded to the nearest five-thousandth, and transformed to z' units (see Tables 2 and 3). Six one-way analyses of variance were performed on these data (three for each of the two rating sessions). Four of these would determine if a rater pair differed significantly on the reliability with which they rated the individuals within a patient group. The remaining two would show differences in rater pairs across groups. No F values approached significance (see Tables 4, 5, 6, 7, 8, and 9) (Edwards, 1960).

In the second analysis of the matrices all coefficients in the monomethod-heterotrait triangles were averaged. These average correlations
Table 2. Reliability for rater pairs on items having non-zero variance, session 2 (decimals have been omitted)

<table>
<thead>
<tr>
<th>Item #</th>
<th>Practice Group rater pairs(^a)</th>
<th>Lecture Group rater pairs(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>1</td>
<td>62</td>
<td>725</td>
</tr>
<tr>
<td>25</td>
<td>70</td>
<td>867</td>
</tr>
<tr>
<td>2</td>
<td>57</td>
<td>655</td>
</tr>
<tr>
<td>7</td>
<td>-17</td>
<td>172</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>177</td>
</tr>
<tr>
<td>24</td>
<td>-01</td>
<td>-010</td>
</tr>
<tr>
<td>9</td>
<td>56</td>
<td>633</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>136</td>
</tr>
<tr>
<td>16</td>
<td>58</td>
<td>655</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>208</td>
</tr>
<tr>
<td>6</td>
<td>39</td>
<td>412</td>
</tr>
<tr>
<td>18</td>
<td>35</td>
<td>371</td>
</tr>
</tbody>
</table>

\(^a\) B = blind; BB = double blind; WW = double wise; W = wise.
Table 2. Reliability for rater pairs on items having non-zero variance, session 3 (decimals have been omitted)

<table>
<thead>
<tr>
<th>Item #</th>
<th>Practice Group rater pairs&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Lecture Group rater pairs&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td></td>
<td>z&lt;sup&gt;1&lt;/sup&gt;</td>
<td>z&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>25</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>24</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>21</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>BB</td>
</tr>
<tr>
<td>18</td>
<td>B</td>
<td>BB</td>
</tr>
</tbody>
</table>

<sup>a</sup>B = blind; BB = double blind; WW = double wise; W = wise.
Table 4. Analysis of variance for rater-pair agreement, rating session 2, Practice group

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>.23</td>
<td>3</td>
<td>.08</td>
<td>1.14</td>
</tr>
<tr>
<td>Within</td>
<td>3.14</td>
<td>44</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3.37</td>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Analysis of variance for rater-pair agreement, rating session 2, Lecture group

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>.06</td>
<td>3</td>
<td>.02</td>
<td>&lt;1.00</td>
</tr>
<tr>
<td>Within</td>
<td>1.43</td>
<td>44</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.49</td>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Analysis of variance for rater-pair agreement, rating session 2, Practice and Lecture groups combined

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>.42</td>
<td>7</td>
<td>.06</td>
<td>1.20</td>
</tr>
<tr>
<td>Within</td>
<td>4.57</td>
<td>88</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4.99</td>
<td>95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7. Analysis of variance for rater-pair agreement, rating session 3, Practice group

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>.09</td>
<td>3</td>
<td>.03</td>
<td>&lt;1.00</td>
</tr>
<tr>
<td>Within</td>
<td>1.67</td>
<td>44</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.76</td>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Analysis of variance for rater-pair agreement, rating session 3, Lecture group

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>.06</td>
<td>3</td>
<td>.02</td>
<td>&lt;1.00</td>
</tr>
<tr>
<td>Within</td>
<td>4.01</td>
<td>44</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4.07</td>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9. Analysis of variance for rater-pair agreement, rating session 3, Practice and Lecture groups combined

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>.27</td>
<td>7</td>
<td>.04</td>
<td>&lt;1.00</td>
</tr>
<tr>
<td>Within</td>
<td>5.68</td>
<td>88</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5.95</td>
<td>95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
were then transformed to z' values and used as data for t-test, i.e., comparisons of the means of the 12 triangles. Data from both the second and third rating sessions were treated in this fashion to determine if the treatment condition to which the patients were assigned could influence the reliability with which the patients were rated. Both sets of data yielded t values less than one. Table 10 contains these values for each item expressed in correlation and z' scale units.

Finally, the correlations for each monomethod-monotrait triangle were averaged to yield a single reliability estimate for that item on each of the two rating sessions. Column 2 of Table 1 provides these estimates and gives an opportunity to compare the reliability estimates when observer bias could be present with the estimates obtained prior to patients being assigned to groups. It can be seen that no large or systematic shifts resulted. In summary, there is no indication that the pairs of raters differ significantly on the reliability with which they use scales either within or between treatment groups.

Rater Preference Within or Between Treatment Groups

From the scale items nine were selected to represent a continuum of reliability. Items 1, 25, and 4 were selected as a high reliability group; items 7, 8, and 24 as a low reliability group; and items 9, 16, and 21 were selected to compose the medium group.

Using these three sets of items as methods and the raters as traits, six 24 x 24 trait-method matrices were constructed. Three such matrices were generated for each rating session; one for each of the two treatment methods for each session (Tables 11, 12, 13, and 14) and one for the
Table 10. Correlations and $z'$ values averaged across raters for items having non-zero variance (decimals have been omitted)

<table>
<thead>
<tr>
<th>Item #</th>
<th>2nd Rating Session</th>
<th>3rd Rating Session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Practice Group</td>
<td>Lecture Group</td>
</tr>
<tr>
<td></td>
<td>$F$</td>
<td>$z'$</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>74</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>63</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>46</td>
<td>63</td>
</tr>
<tr>
<td>16</td>
<td>53</td>
<td>59</td>
</tr>
<tr>
<td>18</td>
<td>31</td>
<td>50</td>
</tr>
<tr>
<td>21</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>24</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>25</td>
<td>75</td>
<td>83</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>-02</td>
<td>-02</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>

combined groups for each session (Tables 15 and 16). Each correlation within the matrix was the correlation between two raters (or a rater and himself).

The first four of these matrices would indicate if the individual raters were showing preferential ratings for patients within a treatment group (Tables 11 to 14). The comparisons made when the treatment groups were combined would indicate a rater who was rating patients from one
Table 11. Trait-method matrix for 2nd rating session. Practice group (decimals have been omitted)

<table>
<thead>
<tr>
<th></th>
<th>High^a</th>
<th>Reliability</th>
<th>Medium^a</th>
<th>Low^a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td>1 2 3 4 5 6 7 8</td>
<td>1 2 3 4 5 6 7 8</td>
<td>1 2 3 4 5 6 7 8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>79 1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>74 94 1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>66 86 87 1</td>
<td></td>
<td>40 46 48 1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>65 74 77 71 1</td>
<td></td>
<td>62 84 64 70 1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>82 89 86 88 73 1</td>
<td></td>
<td>63 64 41 58 75 54 1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>77 91 95 85 73 89 1</td>
<td></td>
<td>87 73 89 86 86 1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>77 91 87 89 75 89 86 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^aHigh reliability = items 1, 25, and 4; medium = items 9, 16, and 21; low = items 7, 8, and 24.

^bRaters 1 and 2 = blind; raters 3 and 4 = double blind; raters 5 and 6 = double wise; raters 7 and 8 = wise.
Table 12. Trait-method matrix for 2nd rating session, Lecture group (decimals have been omitted)

<table>
<thead>
<tr>
<th>Raters</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>1 2 3 4 5 6 7</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>88 1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>88 1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>77 80 1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>81 68 1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>57</td>
<td>86 64 1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>63</td>
<td>80 81 1</td>
<td></td>
</tr>
</tbody>
</table>

Reliability

High \(^a\) = items 1, 25, and 4; medium = items 9, 16, and 21; low = items 7, 8, and 24.

\(^b\) Raters 1 and 2 = blind; raters 3 and 4 = double blind; raters 5 and 6 = double wise; raters 7 and 8 = wise.
Table 13. Trait-method matrix for 3rd rating session, Practice group (decimals have been omitted)

- **Reliability**
  - High
  - Medium
  - Low

<table>
<thead>
<tr>
<th>Raters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8^</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8^</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>89</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>81</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>79</td>
<td>84</td>
<td>60</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>86</td>
<td>75</td>
<td>90</td>
<td>64</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>95</td>
<td>89</td>
<td>94</td>
<td>77</td>
<td>93</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>92</td>
<td>86</td>
<td>91</td>
<td>75</td>
<td>85</td>
<td>93</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td>91</td>
<td>81</td>
<td>78</td>
<td>74</td>
<td>85</td>
<td>88</td>
<td>81</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **High reliability** = items 1, 25, and 4; medium = items 9, 16, and 21; low = items 7, 8, and 24.

- **Raters 1 and 2 = blind; raters 3 and 4 = double blind; raters 5 and 6 = double wise; raters 7 and 8 = wise.**
<table>
<thead>
<tr>
<th>Raters</th>
<th>Higha</th>
<th>Mediuma</th>
<th>Lowa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>290</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>94</td>
<td>87</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>66</td>
<td>57</td>
</tr>
<tr>
<td>5</td>
<td>88</td>
<td>91</td>
<td>88</td>
</tr>
<tr>
<td>6</td>
<td>91</td>
<td>87</td>
<td>91</td>
</tr>
<tr>
<td>7</td>
<td>59</td>
<td>64</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>88</td>
<td>86</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>20</td>
<td>05</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>37</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>-10</td>
<td>01</td>
<td>-07</td>
</tr>
<tr>
<td>4</td>
<td>-02</td>
<td>08</td>
<td>09</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>59</td>
<td>51</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>13</td>
<td>04</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-07</td>
<td>-15</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>-25</td>
<td>-32</td>
<td>-14</td>
</tr>
<tr>
<td>3</td>
<td>-55</td>
<td>-50</td>
<td>-57</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>-61</td>
<td>-49</td>
<td>-48</td>
</tr>
<tr>
<td>6</td>
<td>-06</td>
<td>02</td>
<td>03</td>
</tr>
<tr>
<td>7</td>
<td>-52</td>
<td>-53</td>
<td>-49</td>
</tr>
<tr>
<td>8</td>
<td>-47</td>
<td>-37</td>
<td>-32</td>
</tr>
</tbody>
</table>

ähigh reliabilty = items 1, 25, and 4; medium = items 9, 16, and 21; low = items 7, 8, and 24.

bRaters 1 and 2 = blind; raters 3 and 4 = double blind; raters 5 and 6 = double wise; raters 7 and 8 = wise.
Table 15. Trait-method matrix for 2nd rating session, Practice and Lecture groups combined (decimals have been omitted)

<table>
<thead>
<tr>
<th>Raters 1 2 3 4 5 6 7 8</th>
<th>Reliability</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

^High reliability = items 1, 25, and 4; medium = items 9, 16, and 21; low = items 7, 8, and 24.

^Raters 1 and 2 = blind; raters 3 and 4 = double blind; raters 5 and 6 = double wise; raters 7 and 8 = wise.
Table 16. Trait-method matrix for 3rd rating session, Practice and Lecture groups combined (decimals have been omitted)

<table>
<thead>
<tr>
<th>Reliability</th>
<th>High(^a)</th>
<th>Medium(^b)</th>
<th>Low(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td>1 2 3 4 5 6 7 8(^b)</td>
<td>1 2 3 4 5 6 7 8(^b)</td>
<td>1 2 3 4 5 6 7 8(^b)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>91</td>
<td>84</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>63</td>
<td>74</td>
<td>58</td>
</tr>
<tr>
<td>5</td>
<td>86</td>
<td>82</td>
<td>69</td>
</tr>
<tr>
<td>6</td>
<td>91</td>
<td>86</td>
<td>75</td>
</tr>
<tr>
<td>7</td>
<td>75</td>
<td>74</td>
<td>66</td>
</tr>
<tr>
<td>8</td>
<td>89</td>
<td>83</td>
<td>67</td>
</tr>
<tr>
<td>9</td>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>30</td>
<td>-06</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>19</td>
<td>02</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>19</td>
<td>02</td>
</tr>
<tr>
<td>13</td>
<td>80</td>
<td>89</td>
<td>67</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>01</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>24</td>
<td>27</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>17</td>
<td>-19</td>
<td>-31</td>
<td>-07</td>
</tr>
<tr>
<td>18</td>
<td>-22</td>
<td>-13</td>
<td>-15</td>
</tr>
<tr>
<td>19</td>
<td>-22</td>
<td>-22</td>
<td>-22</td>
</tr>
<tr>
<td>20</td>
<td>-22</td>
<td>-22</td>
<td>-22</td>
</tr>
<tr>
<td>21</td>
<td>-22</td>
<td>-15</td>
<td>-14</td>
</tr>
<tr>
<td>22</td>
<td>-22</td>
<td>-22</td>
<td>-14</td>
</tr>
<tr>
<td>23</td>
<td>24</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>24</td>
<td>46</td>
<td>52</td>
<td>40</td>
</tr>
<tr>
<td>25</td>
<td>55</td>
<td>51</td>
<td>-05</td>
</tr>
<tr>
<td>26</td>
<td>55</td>
<td>51</td>
<td>21</td>
</tr>
<tr>
<td>27</td>
<td>30</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>28</td>
<td>59</td>
<td>82</td>
<td>23</td>
</tr>
<tr>
<td>29</td>
<td>54</td>
<td>37</td>
<td>47</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>42</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>44</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>49</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>51</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>52</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>53</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>54</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>56</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>57</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>58</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>59</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\)High reliability = items 1, 25, and 4; medium = items 9, 16, and 21; low = items 7, 8, and 24.

\(^b\)Raters 1 and 2 = blind; raters 3 and 4 = double blind; raters 5 and 6 = double wise; raters 7 and 8 = wise.
treatment group more favorably than patients from the other (Tables 15 and 16).

These matrices were then analyzed using a procedure based on Campbell and Fiske's method. The requirement from Campbell and Fiske that the values in the validity diagonals be non-zero was dropped as the purposeful use of items having near zero reliability coefficients predetermined that values in the validity diagonal would be near zero. Trait-discrimination would be defined by a value in the validity diagonal exceeding the values within its row and column and also exceeding the values of the other raters within that method.

The data for all matrices clearly indicate that methods (items) and not traits (raters) contributed to the variance, i.e., there is no indication of discriminant validity. Stated simply, the individual raters do not show preferential ratings for either individuals or groups.
DISCUSSION

Five hypotheses were to have been investigated in the present study. They were:

1. Bias will occur when staff members who have worked with experimental groups rate the members of those groups.

2. Bias will be limited to items having low reliability.

3. Bias will not occur when the rater is aware of the possibility of bias but unaware of which individuals are involved in each treatment group; these conditions will lead to a lack of reliability between raters.

4. Bias will be prevented and reliability of the ratings will be preserved when the raters know the reliability of the items employed as well as which subjects are experimental and control.

5. Bias will not be evident in the highly reliable items; these items will not be influenced by any of the conditions of the experiment.

This study has attempted to demonstrate the presence of observer bias in three areas. Conceivably, observer bias would occur when raters 1) differ in the reliability with which they rate members of two treatment groups, 2) use preferential ratings for particular individuals within a single treatment group, and 3) rate individuals from two treatment groups differently, that is, assign suppressed ratings for the nonpreferred group or increase the values assigned to the preferred group.

It was suggested above that raters might differentially rate the members of the two patient groups with respect to the reliability of the ratings. That is, they might rate persons within the preferred treatment group alike and, conversely, rate members from the nonpreferred treatment group as divergent. This might not be apparent in a comparison of the means of the groups but would be reflected by comparing the inter-rater
reliability with which the two groups were rated. This form of bias might or might not affect the means of the groups but would have an effect upon the variance, thereby affecting conclusions based on statistical tests using an estimate of the variance in the denominator. As the F-test performed on the reliability estimates for the rater pairs in the present study failed to demonstrate differences in the reliability of the rater pairs either between or within treatment groups, one must conclude that this form of observer bias was not operating in the present study.

A second way in which raters could bias data would be in biasing their ratings for particular individuals within a treatment group. They could conceivably increase or decrease the means for that group thereby increasing or decreasing the correction factor used in calculating the sum of squares for a test based upon that group. However, if this were done, the correlations between the biasing rater and himself would be greater across methods than the correlations between himself and other raters and should yield a correlation coefficient exceeding the reliability of the item used for the rating. Had this form of bias occurred in the present study, it would have been reflected in the trait-method matrix for each treatment group by showing discriminant validity for the biasing rater. As the matrices showed only method variance, the possibility of this type of biasing having occurred in the present study is ruled out.

The most blatant form which observer bias could take would be one in which a rater systematically increased or decreased his ratings for individuals within his preferred or nonpreferred treatment group. Such a procedure would have obvious effects upon any conclusions drawn about group differences. Had such systematic biasing occurred in the present study,
the trait-method matrices for the combined treatment groups would have shown discriminative validity for those raters who were biasing in this fashion, i.e., they would have tended to systematically assign the persons from their preferred group scores above their mean rating for all items. In this case the raters' correlations with themselves across items would have exceeded the reliability of the nonreliable items and would have exceeded the correlations with other raters both within and between item sets (methods). This did not occur in the present study.

In view of the failure to find bias in the present study, Hypothesis 1, that bias will occur when staff members familiar with the patient assignment to treatments rated the groups, must be rejected. The four raters who had worked with the treatment groups did not bias their ratings by using differential reliability for the patient groups, by assigning preferential ratings to individuals within a single group or by assigning preferential ratings to patients from either treatment group.

Hypothesis 2, that the bias will be limited to items having low reliability, must also be rejected. In the present study data for items with inter-rater reliability as low as .20 were analyzed. If biasing were shown to occur on items below this level, that finding, although interesting, would have little bearing upon the behavior modification controversy. The reliability of the measures customarily applied would exceed this level of reliability.

Had the raters who were familiar with the purpose of the present study but unfamiliar with the patient assignment to treatment group shown less reliability than the other raters, this would have been apparent in the F-test on the reliability coefficients. These raters would have shown
consistent low reliability both within and between treatment groups. There was no indication that there was a difference between raters in the reliability with which they assigned ratings; Hypothesis 3 does not seem tenable.

The fourth hypothesis, which was worded to specify the conditions under which the two individuals who were fully aware of the conditions of the study were rating, cannot be rejected. However, the failure to reject this hypothesis results from the failure to create bias and not upon the specific conditions under which the ratings were made. The knowledge of the reliability of an item apparently has no effect upon the likelihood that the item will be biased if raters who do not have this knowledge also fail to bias the values they assign to patients.

The fifth hypothesis, that high reliability items will show no rater bias, certainly seems acceptable. However, in light of the present study it seems that the definition of high reliability could, so far as being biasable is concerned, be somewhere in the .20 range. Stated bluntly, the reliability of the item cannot be a predictor of the biasability of that item if no items can be shown to be biasable.

In summary, as it was impossible to demonstrate the presence of rater bias in any form, the last four hypotheses become either untestable or irrelevant. It will be recalled that in the Rationale and Procedure chapter of the present study four subgoals were established for the present research. They were: 1) to determine if observer bias would be obtained in a behavioral setting, 2) to attempt a rudimentary scale to estimate the reliability required to prevent bias from occurring, 3) to study the relationship between knowledge of the conditions of the experiment and rater bias, and 4) to demonstrate a technique for assessing observer bias.
In view of the surprising finding that it was not possible to demonstrate observer bias, the first of these goals has been accomplished.

The second goal, that of constructing a scale to estimate the reliability required to prevent observer bias, becomes unnecessary. The raters either assigned zero variance to items having a low reliability, thereby rejecting the item as differentiating between subjects, or used the item without signs of bias. The goal of establishing a scale to estimate the reliability required to prevent bias is thus unnecessary. As mentioned earlier, the raters did not bias items with reliability as low as .20. The likelihood of a researcher in a behavioral setting employing items of less reliability is highly unlikely. The number of items the raters assigned zero variance further suggests if one were to attempt to use less reliable items than those analyzed in this study, the raters would virtually refuse the item by assigning the same score to all subjects.

Despite the occurrence of terms such as "temper tantrums" in the literature, one would probably not be led astray if he concluded that the reliability with which observers could judge the presence or absence of such behavior would either exceed or equal the reliability of some of the items used in the present study or be rejected by the raters.

The final subgoal of the study was to demonstrate a technique for assessing observer bias. The flexibility of the assessment method for dealing with at least three possible forms of bias has been demonstrated.

The failure of the present study to demonstrate observer bias could have occurred as a result of the raters in the key group (W) having sensed that their ratings would be compared to those obtained from the other raters. As the four raters forming the wise and double wise groups were
the only non-Ph.D raters in the group, this could have exerted some unknown effect which might have prevented the bias from occurring. However, had this occurred there is no reason to believe the effect would be limited to this study. In other words, the same effect would be predicted whenever Ph.D. and non-Ph.D. raters were rating the same patients and, if used as a criticism of the present study, would not provide grounds for rejecting any study employing both Ph.D. and non-Ph.D. raters as free of observer bias.

As both the wise and double wise raters in the present study were expecting publication as a result of their work, one could suggest that this prevented the bias by establishing a set or demand that the raters be overly objective. However, this is a description of any experimenter evaluating his work, and, again, if used as a criticism of the present study must not be used as a rationale for criticizing studies where the experimenter expects gain from his work.

Another possibility would be that the rater pairs "guessed" the information that would have given them the same knowledge as raters from another pair. Statements written on the final rating sheet describing the purpose of the ratings indicate that the raters had only the information pertinent to their group and did not have the information which would have placed them in another group.

A final possibility would be that the four raters forming the wise groups were unique in that they were of superior "moral character" and would not, even unintentionally, bias their ratings. If this were the case, future researchers should have little difficulty in demonstrating this.

A final word about observer bias in the form of response sets is appropriate. By referring to the matrices used for determining rater reliability
between patient groups, one may find values on the validity diagonal which show trait-discrimination. This is evidence of a response set and indicates the rater is not differentiating between the people he is rating. Inspection of the two matrices in Appendix III show examples of this.

The primary purpose of the study was to examine the role observer bias played in the behavioral modification controversy. It will be recalled from the review of the literature that the behavior researcher has been criticized on two main points. The first criticism is sampling bias, i.e., the subjects have either been selected because of a specific behavior which fit the behavioral model, have not been selected by an individual other than the experimenter, or have, in the few studies demonstrating sufficient control for sampling, been "normal". Second, there have been criticisms that the literature is open to rater bias because the individuals who were evaluating the subjects were either aware of the treatment the individuals were receiving or had been involved directly with the treatment that was being provided or were members of the client's family or the hospital staff.

It should be clear that in this study the patients being evaluated were: 1) randomly assigned to treatment groups, and 2) composed of chronic mental patients. It should also be clear that there were ratings on these patients by persons who were: 1) hospital personnel, 2) involved with the treatment program, thereby 3) aware of which subjects belonged to the different treatment groups. These ratings were compared with ratings by: 1) nonhospital personnel, who 2) were not involved with the treatment program, thereby 3) not aware of which subjects belonged to the different treatment groups.
From the results of this study it is appropriate to say that, despite the lack of studies safeguarding against the possibility of observer bias, it is doubtful that such bias has exerted influence on the data obtained. A reader in the area of behavior modification need not feel overly concerned by the failure of the behavior researchers to control for observer bias by allowing the experimenter, or a member of his hospital staff, to evaluate the patient.

It would seem advisable that the criticism of behavior modification literature as being plagued with observer bias (Breger and McGaugh, 1965; 1966) be terminated unless or until the critics demonstrate a set of operations capable of eliciting observer bias.
BIBLIOGRAPHY


ACKNOWLEDGMENTS

There are countless people who have been responsible in many ways for providing the opportunity to write this paper. You know who you are and what you have done. Thank you.

I would especially like to acknowledge Drs. Warman, Williamson, Brown and Nord and my graduate committee for their cooperation and support both in this study and in other areas.

Tom Hannum, without the encouragement you gave me many years ago I would not have returned to school. Thank you.

I suppose it is appropriate to say something about the blood, sweat and tears shed by Betty but, in view of a long standing family joke, I will simply say "Despite my wife we finished this".
APPENDIX I: RATING SCALE

RATE THE APPLICANT ON THE FOLLOWING ITEMS ON A RANGE OF 0 TO 9 ACCORDING TO HOW WELL THE ITEMS DESCRIBE THE PERSON. ZERO MEANS THE ITEM DOES NOT DESCRIBE THE APPLICANT AT ALL WHILE NINE MEANS THE ITEM IS HIGHLY DESCRIPTIVE OF THE PERSON

1. Comfortable
2. Relaxed
3. Asks questions of employer
4. Tangential in answering questions
5. Directed himself to raters rather than interviewer
6. Appropriate tone in introduction
7. Initiated introduction before the interviewer did
8. Dressed neatly
9. Answered questions too briefly
10. Did not conclude interview before leaving
11. Avoids reference to his psychiatric stay at hospital
12. Specifies that he has references
13. Volunteers release of information where appropriate
14. Asks for a job
15. Had to be asked if he would like a job
16. Specified what skills he has or does not have
17. Not enough eye-contact
18. Hard to get rid of at the end of the interview
19. Hard to keep his attention
20. Changed subject to non-relevant topic
21. Offered application form right away
22. Had no idea of who to use for reference
23. Inappropriate laughter
24. Makes excuses inappropriately
25. Suggests he can't handle the job before knowing what it entails
26. Excessive body movement
27. Excessive leg movement
28. Hard to hear (interviewer had trouble hearing him)
29. Louder than appropriate
30. Unclear speech even though loud enough
31. Made inappropriate personal comments or questions to interviewer or raters
32. Did not take to actual role or did not understand role
33. Appears motivated
34. Personality appears intact
35. Appropriate affect (emotional expression)
36. High ego strength
37. Tendency to employ magical thinking
38. Expressed paranoid ideas
39. Controlled delusional system during interview
40. Ego integrated
41. Demonstrates thinking disorder
42. Cognitive consistency demonstrated
43. Achievement oriented
44. Regressed
45. High self-concept
46. Reacts to interviewer as a father figure
47. Involved in task
48. Sick talk
49. Finger drumming
50. Stuttering
51. High nervous tension
52. Paranoic looking
53. Reality oriented
APPENDIX II: INSTRUCTIONS

Please do not discuss your ratings or the items with the other raters. Rate the following items according to how well they describe the person in the interview situation. Use the numbers 0 through 9 where 0 means the item is a very poor description of the person and 9 is a very good description of the person. Place your choice on the IBM answer sheets. (Please use a #2 pencil)

Please note: Item 25 should be counted during the interview. The other items should be rated during the interview or immediately after. Ratings should be completed in five minutes from the time the interviewee comes in the room.

1. Dressed neatly
2. Eye-contact
3. Comfortable
4. Hard to hear (interviewer had trouble hearing him)
5. Hard to keep his attention
6. Appropriate affect
7. Nervousness impairs performance
8. Regressed
9. Ego integrated
10. Reacts to interviewer as a father figure
11. Uses sick talk
12. Finger drumming
13. Changed subject to non-relevant topic
14. Inappropriate laughter
15. Excessive leg movement
16. Appears motivated
17. Louder than appropriate
18. Personality appears intact
19. Expressed paranoid ideas
20. Demonstrates thinking disorder
21. High self-concept
22. Makes inappropriate excuses
23. Made inappropriate personal comments or questions
24. High ego strength

25 Number of questions asked ________
APPENDIX III: TWO EXAMPLES OF RESPONSE SET BY RATERS

<table>
<thead>
<tr>
<th>Practice Group</th>
<th>Lecture Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>70  1</td>
<td>1</td>
</tr>
<tr>
<td>45  70  75</td>
<td>1</td>
</tr>
<tr>
<td>80  63  85  47</td>
<td>1</td>
</tr>
<tr>
<td>-36 -46 -30 11</td>
<td>-37 1</td>
</tr>
<tr>
<td>92  77  92 71</td>
<td>-14 1</td>
</tr>
<tr>
<td>84  75  91 67</td>
<td>-37 84 1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>-32 -48 -59 -22 -40 34 -45 -28 1</td>
<td></td>
</tr>
<tr>
<td>-17 -37 -33 -48 16 24 -31 -11 21 1</td>
<td></td>
</tr>
<tr>
<td>22  56 28 -01 01 29 43 28 17 -44 1</td>
<td></td>
</tr>
<tr>
<td>00 41 -03 (78) -07 -79 -14 07 01 -42 11 1</td>
<td></td>
</tr>
<tr>
<td>05 -41 -32 -31 -42 09 -10 -29 52 -33 -05 -01 1</td>
<td></td>
</tr>
<tr>
<td>21 -47 -11 -55 23 -10 -07 08 30 46 -68 -30 45 1</td>
<td></td>
</tr>
<tr>
<td>11 27 07 24 38 -21 00 45 50 30 26 36 -27 06 1</td>
<td></td>
</tr>
<tr>
<td>-29 09 41 20 -63 45 -19 -38 61 -39 71 13 45 -43 06 1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Practice</th>
<th>Lecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>-25 1</td>
<td></td>
</tr>
<tr>
<td>43 54</td>
<td></td>
</tr>
<tr>
<td>-10 73 37</td>
<td>1</td>
</tr>
<tr>
<td>36 04 60 04 1</td>
<td></td>
</tr>
<tr>
<td>03 61 59 03 29 1</td>
<td></td>
</tr>
<tr>
<td>85 -09 68 12 47 -03</td>
<td></td>
</tr>
<tr>
<td>-40 80 11 70 -35 19 -31</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Practice</th>
<th>Lecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>-69 06 -21 -31 03 08 -56 17 1</td>
<td></td>
</tr>
<tr>
<td>-48 17 11 -22 42 28 -28 -07 80 1</td>
<td></td>
</tr>
<tr>
<td>-40 27 -01 13 34 -03 -26 34 74 73 1</td>
<td></td>
</tr>
<tr>
<td>04 43 51 34 21 00 42 31 23 41 51 1</td>
<td></td>
</tr>
<tr>
<td>-49 -03 -15 12 31 -34 -28 20 67 48 77 19 1</td>
<td></td>
</tr>
<tr>
<td>10 -76 -21 -74 32 -27 12 -92 19 39 -03 -12 08 1</td>
<td></td>
</tr>
<tr>
<td>-41 -13 02 -16 13 -14 -16 09 66 34 41 12 76 13</td>
<td></td>
</tr>
<tr>
<td>-31 38 00 15 01 10 -28 (66) 64 32 77 36 60 -45 50 1</td>
<td></td>
</tr>
</tbody>
</table>