J. Psychiat. Res. Vol. 2, pp. 61–72. Pergamon Press Ltd. Printed in Great Britain

A LONGITUDINAL STUDY OF THE EFFECTS OF EXPERIMENTER BIAS ON THE OPERANT LEARNING OF LABORATORY RATS*

ROBERT ROSENTHAL and REED LAWSON

Dept. of Social Relations, Harvard University, Cambridge, Mass.
and
Dept. of Psychology, Ohio State University, Columbus, Ohio.

(*Received* 20 *August* 1963)
(*Revised* 4 *December* 1963)

IT NOW appears to be a well-established finding that under a number of differing conditions, psychological experimenters tend to obtain from their human subjects the data experimenters expect or want to obtain.[1,2,3,4,5] In addition, one study has been carried out which showed that the experimenter bias phenomenon may also occur when the subjects (*S*s) are laboratory rats.[6] In that study, twelve experimenters (*E*s) each ran five albino rats on a simple discrimination problem, daily for a five day period. Half of the *E*s were told that the rats they were running had been bred for maze-brightness while the remaining *E*s were told that their *S*s had been bred for maze-dullness. The animals actually assigned to each *E* were standard laboratory animals randomly assigned to *E*. Results of this study clearly indicated that those *E*s believing their *S*s to be maze-bright obtained significantly better performance from their *S*s than did the *E*s believing their *S*s to be maze-dull.

The purpose of the present experiment was to test the generality of these findings by studying the effects of *E*s' biases on the performances of rats in an extended series of Skinner box problems.

METHOD

Experimenters

The 30 male and 9 female students enrolled in a laboratory course in experimental psychology served as *E*s. At the very beginning of the course, all *E*s were given the following written instructions:

"The reason for running these experiments is to give you experience in duplicating experimental findings and, in addition, to introduce you to the field of animal research and overcome any fears you might have with regard to working with rats.

*The research program of which this study is a portion has been supported by research grants G17685 and G24826 from the Division of Social Sciences of the National Science Foundation. This study was conducted while RR was also at the Ohio State University.

The experiments are all repetitions of work done recently on Skinner box-Bright and Skinner box-Dull rats. Many studies have shown that continuous inbreeding of rats that do well on Skinner box problems, such as those you will be running, leads to successive generations of rats that do considerably better than 'normal' rats. Furthermore, these studies have shown that continuous inbreeding of rats that do badly on Skinner box problems, such as those you will be running, leads to successive generations of rats that do considerably worse than 'normal' rats.

Thus, generations of Skinner box-Bright rats do much better than generations of Skinner box-Dull rats.

Each of you will be assigned to a group to work with. Some groups will be working with Skinner box-Bright rats, others will be working with Skinner box-Dull rats.

Those of you who are assigned the Skinner box-Bright rats should find your animals on the average showing some evidence of learning during even the early stages of each of your experiments. Thereafter, performance on each of your experiments should rapidly increase.

Those of you who are assigned the Skinner box-Dull rats should find on the average very little evidence of learning in your rats. You should, however, not become discouraged since it has been found that even the dullest rats can, in time, learn the required responses.

If you are interested in learning more about the details of the experiments on breeding rats for brightness and dullness, your lab instructors can give you references to the work done by Tryon and others at the University of California at Berkeley and elsewhere".

Subjects

Sixteen female laboratory rats (all 80 days old) drawn from the animal colony maintained by The Ohio State University Department of Psychology were randomly assigned to one of two groups. One group of eight *S*s was assigned to home cages which had been labelled 'Skinner-Box Bright' while the other group was assigned to home cages which bore the labels 'Skinner-Box Dull'. Early in the course, two of the *S*s labelled dull died so that the maximum N for the subsequent experiments was eight *S*s labelled 'bright' and six *S*s labelled 'dull'. All *S*s were on a feeding regimen of 1/2 hr *ad lib* access to food daily throughout the 8 wk of the study.

Materials

The basic equipment employed in the studies were commercially made (Scientific Prototype Co.) demonstration Skinner boxes with feeders that dispensed 45-mg. P. J. NOYES pellets.

*E*s followed the laboratory manual of HOMME and KLAUS[7] except that food pellets were used instead of water as reinforcement.

A questionnaire consisting of a series of 30 (20-point) rating scales on which *E*s could rate their satisfaction with their participation in the experiments, their feelings about their *S*s, and their perception of their own behavior during the conduct of the experiments was administered at the conclusion of the study. Each of the scales ran from minus 10 (e.g., extremely dissastisfied) to +10 (e.g., extremely satisfied) with intermediate labelled points.

Space was also provided on these questionnaires for each *E* to describe in his own words how he felt about his participation in these experiments.

Procedure

At the beginning of the study, each *E* was assigned to one of five laboratory periods, each of which had been assigned one or two 'bright' and one or two 'dull' *S*s. Assignment of *E*s to laboratory sections could not be random since there were only certain times that certain *E*s were able to schedule their laboratory section. Within each laboratory section, however, *E*s were randomly assigned to the *S*s to be run during that laboratory section. At least two *E*s were assigned to each *S* and the mean number of *E*s per *S* was 2·7.

Each laboratory team performed three different functions during each of the experiments; that of experimenter, timer and recorder. These functions were rotated among the *E*s comprising each laboratory team. For those teams consisting of only two members, the function of timer and recorder were usually performed by the same *E*.

A total of seven experiments were performed, each of which is described in the manual mentioned earlier.[7] A brief description of each follows:

Experiment I magazine training. Training the rat to run to the magazine and eat whenever the feeder was clicked. Latencies were recorded for each click and the dependent variable was defined as the mean of the mean latencies on the first and of the last ten clicks of the session.

Experiment II operant acquisition. Training the rat to bar-press. Number of bar-pressing responses per min was recorded and the dependent variable was defined as the mean of the mean number of responses during the first and of the last ten minutes of the session.

Experiment III extinction and spontaneous recovery. Number of responses per min was again recorded and the dependent variable was defined as the number of minutes elapsed until the animal showed two response-free minutes. Data were analyzed separately for extinction and for spontaneous recovery.

Experiment IV secondary reinforcement. In this experiment, the animals' responses were reconditioned and partially re-extinguished. Subsequent responses were reinforced by the clicking sound without presentation of food. The dependent variable was again defined as the number of minutes elapsed until the animal showed two response-free minutes, while getting click reinforcements.

Experiment V stimulus discrimination. Training the rat to bar-press only in the presence of a light and not in the absence of the light. For each trial of this experiment, *E*s recorded the latency for the reinforced response and the number of responses occurring under the non-reinforced condition until a criterion of 30 sec of no responding had been reached. The dependent variable was defined as the ranks of the mean latencies of the first and last ten trials added to the ranks of the mean number of non-reinforced responses during these same trials.

Experiment VI stimulus generalization. Demonstrating that animals trained to respond only in the presence of a 110V light would show a decrease in response rate as the voltage

was decreased to 70V to 35V, and finally to 0V. For each of the four test periods, the number of responses was recorded and the dependent variable was defined as the probability for each *S* that his response decrements as a function of stimulus decrements could have occurred by chance. The ranks of these probabilities would, of course, be identical or nearly so, to the ranks of any other index of monotonic decrease.

Experiment VII chaining of responses. Conditioning a loop-pulling response which was followed by the light which signalled the animal that a bar-press would produce a food pellet. The number of complete chains per min was recorded and the dependent variable was defined as the mean number of completed chains during the first and last 10 min.

The students were expected to complete each of these studies in one 2 hr period each week, excepting the stimulus discrimination which was given two periods to complete. If a team did not complete a study within the scheduled time, they had to return to the laboratory in their free time and continue working until *S* was ready to go on to the next scheduled experiment.

Comparison with earlier study using animal Ss

There are several differences between this and the ROSENTHAL-FODE[6] study. The studies were done at different universities using different learning tasks and apparatus. In this study, there were fewer *S*s, 14 compared to 60, but more *E*s, 38 instead of 12. In addition, this was a longitudinal study lasting about 8 wk and a minimum of 14 hr spent with each *S*, while the earlier study lasted 1 wk (5 hr). In the present study, in spite of *E*s rotating their team functions, each *E* spent a minimum average of 4 hr working with his *S*, while in the earlier study no *E* spent more than 1 hr with any one of his five *S*s.

In the earlier study, *E*s worked alone and were much of the time unobserved by the laboratory supervisor. While those instances of cheating which came to light were found to be randomly distributed over the two treatment conditions, the present study provided better control over this possibility since a laboratory instructor was present during each of the laboratory periods.* Perhaps more important than the control of cheating was the control of gross cues to *S*s. Thus if an *E*, because of his belief that a rat was dull, handled the animal roughly, the laboratory instructor was there to point out to the *E* that *S* would never learn unless he were better treated. In the present study too, the motivations of *E*s were quite different. In the earlier study it was found that *E*s felt better when their *S*s learned well, but there was no external sanction for their learning well. In the present situation, the rat in effect *had* to learn in order that *E* could write a report, get a grade and go on to the next study. An additional motivational difference was possibly associated with the differing roles of the laboratory instructors in the earlier and the present study. In the earlier study, the lone laboratory instructor reinforced *E*s' beliefs that poor performance was accounted for by the rats' dullness. In the present study, only one of the three instructors did so. Another instructor evaded any reference to the rats' brightness or dullness, while the third instructor told his *E*s that there was no such thing in the final analysis as a

*The laboratory instructors, in addition to LAWSON, included DONALD J. BARNES and WILL K. WEINSTEIN, whose cooperation is gratefully acknowledged.

dull rat; only dull *E*s! Fortunately, then, we have acquired a small sample of 'climates' apparently more or less favorable to the occurence of experimenter bias, thus increasing somewhat the generality of any findings.

RESULTS

Preliminary inspection of the data revealed that for several of the experiments there were such extremely deviant scores that the use of interval scale statistics seemed inappropriate. Therefore, in each experiment, the obtained scores were converted to ranks and the treatment effect evaluated by means of the Mann-Whitney *U* test. Since on the average each experiment was not conducted by some team of *E*s from each treatment condition, the mean raw rank for each treatment group was not comparable from experiment to experiment. In order to achieve comparability of mean ranks across experiments, and to legitimize their addition, all ranks were converted to GUILFORD'S C-scale scores.[8]

For each *S* a mean rank based on performance in all experiments was computed. The performance of those *S*s who had been labelled 'bright' was significantly superior to the

TABLE 1. NORMALIZED MEAN RANKS OF OPERANT LEARNING FOR EACH EXPERIMENT (LOWER RANKS INDICATE SUPERIOR LEARNING)

	Experiment	(N)	Bright	Dull	(N)	Mann-Whitney *U* One-tailed *p*	Rank Correlation with preceding experiment
I.	Magazine Training	(7)	4·4	5·8	(5)	0·13	
II.	Operant Acquisition	(8)	4·3	6·2	(5)	0·09	0.25
III. A.	Extinction	(6)	4·2	5·8	(5)	0·12	0·08
B.	Spontaneous Recovery	(8)	4·6	5·0	(6)	0·48	0·25
IV.	Secondary Reinforcement	(6)	4·7	5·5	(4)	0·17	0·37
V.	Stimulus Discrimination	(8)	4·0	6·3	(6)	0·008	0·38
VI.	Stimulus Generalization	(7)	4·3	5·8	(4)	0·02	0·59 ($p < 0·05$, one-tailed)
VII.	Response Chaining	(8)	5·8	3·8	(5)	0·17 (two-tailed)	0·45
	Means	(7·3)	4·5	5·5	(5·0)	0·015	0·35 ($p < 0·005$, one-tailed)

performance of *S*s labelled 'dull' ($p = 0{\cdot}015$).* Table 1 shows the normalized mean ranks for each treament group as well as the Spearman rank correlation of the performances in each experiment with the performances of the preceding experiment. Inspection of the p levels obtained for the eight comparisons reveals seven probabilities of 0·17 or less. Of these seven, 1·3 would be expected by chance in the computation of eight *U*s. It appears likely then that *E*s' belief or expectation about the performance of their rat *S*s affected the performance obtained from their *S*s in most of the experiments. One of the experiments, the final one on response chaining, showed a tendency to give significant results in the nonpredicted direction. Inspection of the p levels for the eight comparisons suggested no trend for subsequent treatment effects to become either more or less significant. The median p level for the first four comparisons was 0·13 and for the last four comparisons it was 0·10.

The question of correlated performances must now be faced. That is, did the differences between the treatment groups arise during the first experiment and then simply maintain themselves over subsequent experiments? The answer to this question will tell us in part whether our seven experiments were nothing more than a single experiment replicated seven times. In Table 1, the last column shows that in most cases less than 15 per cent of the variance of performances in any experiment could be accounted for by the variance of performances in the preceding experiment. Only the correlation between performances in the experiments on stimulus discrimination and generalization was significant at the 5 per cent level, one-tailed test. A good illustration that the amount of correlated variance was not a crucial factor is provided by examination of the results of the Response Chaining experiment in Table 1. In that experiment, *S*s' performances correlated 0·45 with their performances in the preceding experiment, this correlation accounting for about 20 per cent of the variance. Yet in spite of this, the obtained mean performances differed significantly from each other in the opposite directions! Nevertheless, we must conclude that overall performances did tend to be statistically significantly correlated, since the mean correlation computed using Fisher's z-transformation was 0·35 which with pooled df was significant at the 0·005 level, one-tailed test. In addition, inspection of the correlations for succeeding experiments suggests a trend for them to get larger.

The original assignment of *S*s to treatment conditions had been random, but the question may fairly be asked whether by chance the *S*s labelled 'bright' might not in fact have been brighter, especially in view of the small sample size.† This question cannot be answered directly, but the likelihood of this factor accounting for our overall results can be evaluated. If the obtained results had been due to pre-experimental differences among the *S*s rather than to the labelling treatment, we would have expected correlations differing significantly from zero between *S*'s performance in an experiment and her performance in the subsequent experiment. As an additional check on this question, the following comparison was made. Those four 'dull' *S*s who participated in both experiments I and II were matched with those four 'bright' *S*s who also performed in both experiments and whose performances in experiment I were most similar to that of the 'dull' *S*s. The mean normalized rank of per-

*The p levels given are based on one-sided tests since the direction of difference was predicted. Following strictly the logic of one-sided tests would not permit us to consider the significance of the nonpredicted result of experiment VII for which the two-tailed p has, nevertheless, been given.

†Thanks are due MAX BERSHAD and LEON PRITZKER for pointing this problem out.

formance in experiment I for the four 'dull' *S*s was 5·5 while for the four 'bright' *S*s the mean was 5·8. The mean normalized rank of performance in experiment II was 6·0 for these 'dull' *S*s and 3·8 for these 'bright' *S*s, the difference being significant at the 0·10 level, one-tailed *U* test. Thus, when we considered experiment I as the basis for pre-experimental matching, the subsequent experiment showed no change in the direction of difference or the degree of its significance, from what was obtained without pre-experimental matching. It seems reasonable then, to assume that if pre-experimental differences in ability favored the 'bright' *S*s, this could at most only have affected the results of experiment I.

It was mentioned earlier that all *S*s had been assigned to one of five laboratory periods, one or more of which was supervised by one of three laboratory instructors. It was also mentioned that each of these instructors appeared to provide a somewhat different 'climate' which might be interepreted as more or less favourable to the occurrence of experimenter bias. Table 2 shows the normalized mean ranks for all the experiments combined for each

TABLE 2. NORMALIZED MEAN RANKS OF OPERANT LEARNING FOR EACH LABORATORY (LOWER RANKS INDICATE SUPERIOR LEARNING)

Laboratory	(N)	Bright	Dull	(N)	Mann-Whitney *U* One-tailed *p*
A	(7)	4·3	5·3	(14)	0·08
B	(15)	4·9	6·5	(6)	0·07
C	(15)	5·1	5·8	(8)	0·25
D	(7)	3·7	4·6	(7)	0·21
E	(14)	4·1	6·0	(5)	0·07
Means Total	(11·6)	4·4	5·6	(8·0)	0·02

treatment group listed by laboratories. The N indicates, for each treatment in each laboratory, the number of experiments represented. An N greater than eight, of course, means that there were two *S*s run in that treatment condition in that laboratory. In all 5 laboratories, the treatment effects were in the predicted direction with *p* levels ranging from 0·07 to 0·25. The differences in obtained *p* levels for such a small sample of laboratories do not seem to warrant elaborate interpretation, although it seems safe to say that 'climates' or lab periods did not seem to make much difference.

At the conclusion of the experiments, each *E* filled out the questionnaire described earlier. Table 3 shows the mean rating on each of the thirty scales for both treatment groups. Those scales listed under Clusters I, II and IV have been grouped together because of thier common membership in what earlier research has shown to be a significant cluster.[2] With the exception of those scales labelled as 'new' in Table 3, all of the scales were employed in the earlier study of experimenter bias using animal *S*s.[6]

TABLE 3. MEAN RATINGS OF *Ss* AND SELF

	Bright	Dull	t	p <0·10 (two-tailed)
I. Satisfaction with Experiment	9·1	6·6	4·40	0·001
II. Ratings of *Ss*				
1. Aggressive (new scale)	3·3	1·2	<1	
2. Healthy (new scale)	6·5	6·1	<1	
3. Friendly (new scale)	2·5	5·5	1·32	
4. Bright	5·0	−2·6	2·64	0·03
5. Clean	5·9	6·1	<1	*
6. Tame	2·6	4·5	<1	*
7. Pleasant	5·2	3·7	1·07	
8. Like	4·1	2·2	1·28	
III. Self Ratings				
A. Cluster I				
1. Honest	7·9	6·5	1·09	
2. Relaxed	6·3	6·1	<1	
3. Casual	2·8	4·7	1·10	*
4. Business-like	3·4	4·6	<1	
5. Pleasant-voiced	4·7	6·0	<1	*
6. Behaved consistently	4·7	6·6	<1	*
7. Pleasant	6·4	5·3	2·04	0·07
B. Cluster II				
1. Friendly	5·8	3·9	1·31	
2. Interested	6·6	7·1	<1	*
3. Encouraging	4·3	3·3	<1	
4. Personal	3·2	6·0	1·56	
C. Cluster IV				
1. Non-talkative	−0·2	−2·7	1·11	
2. Enthusiastic	4·2	1·9	1·31	
3. Professional	2·2	2·9	<1	*
D. Non-loud	3·4	1·5	1·01	
E. Gentle Handling of *Ss*	5·8	5·7	<1	
F. Much Handling of *Ss*				
1. Before each experiment	−1·2	−2·3	<1	
2. After each experiment	−0·8	−3·3	1·17	
[3. Total Handling	−1·0	−2·8	3·34	0·09]
G. Much Watching of *Ss* (new scale)	9·3	8·3	2·16	0·06
H. Much Talking to *Ss* (new scale)	−3·1	0·7	2·41	0·04
I. Relative Performance of *Ss* (new scale)	8·2	−2·9	4·51	0·001

*Indicates opposite direction of mean difference from earlier study.

In the present study, *E*s believing their *S*s to be bright were significantly more satisfied with their participation in the experiments than those *E*s believing their *S*s to be dull. However, even these latter *E*s were remarkably satisfied (6·6) compared to either the *E*s running 'bright' *S*s (3·0) or those running 'dull' *S*s (2·5) in the earlier study, although even then the difference tended to be significant (two-tailed $p = 0{\cdot}10$). This much greater satisfaction of *E*s in the latter study may have been due in part to the nature of the experiments performed by *E*s. These were an integral part of the course content and the principles they were designed to demonstrate were covered in lectures by the course instructor. In the earlier study, there was relatively much less relationship of the experiment to the content of the course. Furthermore, in the earlier study the 'bright' rats learned faster but none really learned well, while in the present study, although the 'bright' rats again learned faster, almost all animals did learn eventually.

That *E*s running the 'bright' rats rated these animals as superior to those they considered dull comes as no surprise and requires no comment. Other differences in the ratings suggest that in this study, those *E*s believing their *S*s to be bright saw themselves as more pleasant in relation to their *S*s, as handling them more, watching them more but talking to them *less*. These differences were significant at beyond the 0·10 level, two-tailed test. The first two of the last four mentioned scales showed the same direction of differences in the earlier study and at about the same levels of significance. The last two mentioned scales were both new to this study.

No other scales reached the 0·10 level, two-tailed test in *both* studies. There were a number of scales, however, which in both studies tended toward significance. In both studies, *E*s believing their *S*s to be bright rated themselves as more enthusiastic, encouraging, and more non-loud, friendly, less personal, less business-like and less talkative in their relationship with their *S*s. In addition, these *E*s liked their *S*s better and found them more pleasant to work with. For these last mentioned nine scales, the mean difference in ratings for the two treatment groups for both studies exceeded either 2·7 scale units or a two-tailed p level of 0·20 in each case.

The last item on each questionnaire was an open-ended question asking each *E* to say in his own words how he felt about the experiments. Nineteen completed questionnaires were obtained from *E*s who had worked with 'bright' *S*s and 17 were received from *E*s who had worked with 'dull' *S*s. Of the former group, 63 per cent stressed the benefit they derived from the experiments, 53 per cent stressed their interest in the experiments, 5 per cent stressed difficulties in *S*'s learning and 0 per cent failed to write any comments. Of the group working with 'dull' *S*s, 41 per cent stressed the derived benefits, 18 per cent stressed their interest, 47 per cent stressed difficulties in *S*'s learning and 12 per cent failed to write any comments. One comment was found to be especially interesting: "Our rat, number X, was in my opinion, extremely dull. This was especially evident during training for discrimination. Perhaps this might have been discouraging but it was not. In fact, our rat had the 'honor' of being the dullest in all the sections. I think that this may have kept our spirits up because of the interest . . . in (our) rat". As a matter of fact, the animal in question was one of the two animals performing at the median level on the discrimination problem as well as for all the experiments taken as a whole. The cited comment serves to point out anecdotally the importance to the *E*s of the type of rat they were running. None of the 34

written comments even remotely suggested that any *E* was aware that their *S*s had not been specially bred. The three laboratory instructors confirmed this lack of suspicion on the part of the *E*s.

On the last day of the course, after the experiments and the questionnaires had been completed, the writers explained the entire study to all the *E*s. There appeared to be great interest and animation on the part of the *E*s. One reaction, though, was outstanding; the unwillingness of many of the *E*s whose rats were 'dull' to believe that their *S*s had not really been bred for dullness, and that in fact they often originated from the same litters as the 'bright' rats.

DISCUSSION

It now appears quite safe to conclude that experimenter bias can effect a given outcome in research studies using animal *S*s and student *E*s. It is also apparent that the effect of this bias is marked enough to be clearly demonstrable with an N of only 14 *S*s. And yet, while the effect itself is marked, its mode of mediation from *E* to *S* may be very subtle indeed. It appears unlikely that gross differences in the treatment and handling of *S*s (or in the recording of data) would have gone unnoticed and uncorrected by the various laboratory instructors whose task it was to supervise the learning of the *E*s via the learning of the *S*s. Also relevant to the robustness of the phenomenon was the fact that because of their primary teaching function, the laboratory instructors tended to give more help and advice to the *E*s whose *S*s were performing more poorly, a fact which would tend to decrease obtained treatment effects.

What can be said specifically about the several experiments showing greater or lesser experimenter bias? Are certain types of tasks which rats may be called upon to perform more susceptible to experimenter bias? It seems doubtful that our data can answer this question. We may feel most confident in *E*s' ability to obtain biased data on stimulus discrimination and generalization type experiments. However, it might prove most useful for the present at least, to regard our median obtained *p* level of 0·13 as our best estimate of the median *p* level to be obtained, with similar sample sizes, if we were to continue sampling the population of operant learning experiments. Taking this view, our more extreme *p* levels, those closer to zero and closer to one, would be regarded as sampling fluctuations.

Before discussing the findings from the questionnaire data analysis, we must ask about the accuracy of this type of data. When an *E* says he was friendly, more or less, was he actually friendly, more or less? Earlier studies[2] utilizing most of the rating scales employed in the present study suggest that the scales may well be reasonably accurate. In one study, 12 *E*s rated their own behavior during the experiment and their *S*s also rated them at the conclusion of the experiment. The correlation (rho) between *S*s' mean ratings of *E*s and *E*s' mean ratings of themselves was 0·89, which with 25 *df* was significant at well beyond the 0·001 level.

Our next question concerning *E*s' ratings of their behavior toward their *S*s is whether it would make more sense to regard these ratings and the behaviors they might portray as antecedents or consequents of the performances of *S*. Perhaps it would make most sense to regard them as both. Thus, initially, those *E*s expecting their *S*s to perform in dull

fashion treated their *S*s in some subtle fashion such as to produce dull behavior while those *E*s expecting bright performance treated their *S*s accordingly. These initial differences in the treatments accorded *S*s might lead to different performances by *S*s which would, in turn, reinforce *E*s' expectations about their *S*s and maintain the subtle treatment differences.

Modally, what was the behavior of the *E* who thought his *S* was bright compared to the behavior of the *E* who thought his *S* was dull? The former tended to be more enthusiastic, friendly, encouraging, pleasant, and more interested in *S*'s performance. He liked *S* more, watched him more intently, and found him to be more pleasant himself. He talked less to his *S* and he was less loud around him, leading one to wonder just what *E*, believing *S* to be dull, might have been saying to him either out loud or perhaps under his breath! Perhaps the most important difference between our modal *E*s was in the amount of handling. Brighter *S*s were apparently handled more, or at least their *E*s thought they were handling them more, which might reflect qualitative rather than quantitative handling differences. BERNSTEIN[9] found that rats handled *more* learned better, and CHRISTIE[10] and others have even been able to tell which *E* had handled an *S* by observing *S*'s behavior in a maze or while being picked up. Our best hypothesis accounting for the results of the present study and based on questionnaire data would seem to be that more handling and less talking and less loudness accounted for the observed differences in performance. Whether rats are sensitive to attitudinal differences in *E* other than through tactual and auditory sense modalities remains to be seen on the basis of further research. As a matter of fact, those aspects of tactual and auditory stimulation that may make the subtle but crucial difference are also unknown at the present time.

SUMMARY

A total of 38 *E*s were divided into 14 research teams, each of which had one rat assigned to it. Eight of the teams were told that the *S*s they would be working with had been bred for brightness while the remaining six teams were told that their *S*s had been bred for dullness. All *S*s were drawn from the same animal colony, all were female and all were 80 days old. All *S*s were assigned at random to one of the two treatment conditions which were *E*s' beliefs or expectations about *S*s' ability.

Seven experiments including (I) magazine training, (II) operant acquisition, (III) extinction and spontaneous recovery, (IV) secondary reinforcement, (V) stimulus discrimination, (VI) stimulus generalization, and (VII) chaining of responses, were performed. Differences in performance favored the groups of *E*s believing their *S*s to be bright in 7 out of the 8 comparisons (overall $p = 0{\cdot}02$). There was no trend over the course of the experiments for the treatment effects to either increase or decrease nor were the performances of *S*s in any experiment save one correlated significantly with their performances in the subsequent experiment. It appeared, then, that the several experiments were, to a great extent, independent. Comparisons of the treatment effects among each of the 5 laboratory sections to which *E*s had been assigned showed no real difference, all sections showing the mean differences in the predicted direction and at similar levels of significance.

These differences were obtained in spite of the fact that laboratory instructors gave more help to *E*s whose *S*s were performing poorly, and that all *E*s were motivated to have their

*S*s perform well in order to complete the sequence of experiments. In addition, a laboratory instructor was present in each laboratory so that gross differences in *E*s' experimental procedure and treatment of their *S*s would have been observed and corrected.

On the basis of questionnaire data obtained in this and in an earlier study, it appeared that *E*s believing their *S*s to have been bred for brightness were more satisfied with their participation in the experiments, liked their *S*s more, watched them more intently and found them to be more pleasant. They tended also to be more enthusiastic, friendly, encouraging, pleasant and interested in their rat's performance, but were less talkative and less loud when working with their *S*. But perhaps the crucial difference was that these *E*s may have handled their *S*s more; a difference which could, on the basis of other research,[9] account for their superior learning.

REFERENCES

1. ROSENTHAL, R. Experimenter outcome-orientation and the results of the psychological experiment. *Psychol. Bull.*, (in press).
2. ROSENTHAL, R. On the social psychology of the psychological experiment: the experimenter's hypothesis as unintended determinant of experimental results. *Amer. Sci.*, 1963, **51,** 268–283.
3. ROSENTHAL, R. and FODE, K. L. Three experiments in experimenter bias. *Psychol. Rep. Monog.*, **3, 12,** 1963.
4. ROSENTHAL, R., PERSINGER, G. W., VIKAN-KLINE, LINDA, and FODE, K. L. The effect of experimenter outcome-bias and subject set on awareness in verbal conditioning experiments. *J. verb. Learn verb. Behav.*, 1963, **2,** 275–283.
5. ROSENTHAL, R., PERSINGER, G. W., VIKAN-KLINE, LINDA, and MULRY, R. C. The role of the research assistant in the mediation of experimenter bias. *J. Personality*, 1963, **31,** 313–335.
6. ROSENTHAL, R. and FODE, K. L. The effect of experimenter bias on the performance of the albino rat. *Behav. Sci.*, 1963, **8,** 183–189.
7. HOMME, L. E. and KLAUS, D. J. *Laboratory studies in the analysis of behavior.* Pittsburgh: Lever Press, 1957.
8. GUILFORD, J. P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.
9. BERNSTEIN, L. The effects of variations in handling upon learning and retention. *J. comp. physiol. Psychol.*, 1957, **50,** 162–167.
10. CHRISTIE, R. Experimental naivete and experiential naivete. *Psychol. Bull.*, 1951, **48,** 327–339.