That an experimenter can very easily influence his subjects to give him the response he wants is a problem that every investigator recognizes and takes precautions to avoid. But how does one cope with the problem of unconscious influence? It is possible that a good many contradictory or unexpected findings are due to the fact that the experimenter unknowingly communicated his desires or expectations to his subjects. Though this problem has been generally recognized and much discussed, there has heretofore been no systematic test of the hypothesis that an experimenter can obtain from his subjects the data he expects or wants to obtain. This paper reports just such a test, and discusses what can be done to avoid experimenter influence, both conscious and unconscious.

THE EFFECT OF EXPERIMENTER BIAS ON THE PERFORMANCE OF THE ALBINO RAT

by Robert Rosenthal and Kermit L. Fode

Harvard University and University of North Dakota[1]

RECENT studies in experimenter bias have shown that *E*s are able to obtain from their human *S*s the data that *E*s expect or want to obtain (Fode, 1960; Rosenthal & Fode, 1961; Rosenthal, Fode, Friedman, & Vikan, 1960; Rosenthal, Fode, & Vikan, 1960; Rosenthal, Friedman, Johnson, Fode, Schill, White, & Vikan, 1960). The general importance to the outcome of experiments of *E* effects has been much discussed in the social and clinical psychological research literature and has been reviewed elsewhere (Rosenthal & Fode, 1961).

References to *E* effects can also be found in the literature of "experimental" psychology and to a much greater extent in informal communications among experimental psychologists. Fattu and Mech (1953, p. 154) state: "That experimenters in the area of learning often are likely to underestimate their roles in the situation is known well." Maier (1956) has more recently discussed this issue and related an anecdote suggesting that an *E* characteristic may have influenced the outcome of an experiment using rat *S*s. Most recently in an exchange of letters in *Science*, Zirkle (1958) and Razran (1959) in discussing Pavlov's attitude toward the notion of the inheritance of acquired characteristics give credence to a statement by Gruenberg (1929, p. 327): "In an informal statement made at the time of the Thirteenth International Physiological Congress, Boston, August 1929, Pavlov explained that in checking up these experiments it was found that the apparent improvement in the ability to learn, on the part of successive generations of mice, was really due to an improvement in the ability to teach, on the part of the experimenter! And so this 'proof' of the transmission of modifications drops out of the picture, at least for the present." It would appear then that Pavlov was indeed aware of the possibility of *E* influencing animal *S*s. With all of this discussion of *E* effects on animal *S*s, there has been no systematic attempt to demonstrate either the occurrence of this phenomenon or its reliability. The purpose of the present study was to test the hypothesis that *E*s are able to obtain from their animal *S*s the data they want or expect to obtain.

METHOD

Experimenters

Twelve of the thirteen students enrolled in a senior division course in experimental psychology served as *E*s. All had been performing experiments as part of the course during the

[1] We want to thank Dr. Ralph H. Kolstoe for his advice regarding animal handling procedures, and Miss Linda L. Vikan and Mr. Gordon Persinger, National Science Foundation Undergraduate Research Participants, for their assistance in the conduct of this experiment and in the analysis of the results.

entire semester and the present study was assigned as their last experiment. All *E*s were given the following written instructions:

Instructions to Es: "The reason for running this experiment is to give you further experience in duplicating experimental findings and, in addition, to introduce you to the field of animal research and overcome any fears that you may have with regard to working with rats.

"This experiment is a repetition of work done on maze-bright and maze-dull rats. Many studies have shown that continuous inbreeding of rats that do well on a maze 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 a maze leads to successive generations of rats that do considerably worse than 'normal' rats.

"Thus, generations of maze-bright rats do much better than generations of maze-dull rats.

"Each of you will be assigned a group of five rats to work with. Some of you will be working with maze-bright rats, others will be working with maze-dull rats.

"Those of you who are assigned the maze-bright rats should find your animals on the average showing some evidence of learning during the first day of running. Thereafter performance should rapidly increase.

"Those of you who are assigned the maze-dull rats should find on the average very little evidence of learning in your rats.

"The experiment itself will involve a discrimination-learning problem. The animals will be rewarded only if they go to the darker of two platforms. In order that the animals do not simply learn a position response, the position of the darker platform will be varied throughout each day's running."

Subjects

A total of 65 naive, Sprague-Dawley albino rats ranging in age from 64 to 105 days were divided into thirteen groups of five each, in such a way as to minimize differences in mean age per group. Each group was comprised of two male and three female *S*s and ranged in mean age from 83 to 91 days. Each group was housed in two cages segregated by sex of *S* and several days before the beginning of the experiment placed on 23-hour food deprivation.

A simple elevated T maze described by Ehrenfreund (1952) was constructed to his specifications. The two arms were interchangeable and one was painted white while the other was painted dark gray.

A questionnaire similar to one used in earlier studies (Rosenthal, Fode, Friedman, & Vikan, 1960) was constructed which consisted of a series of 20-point rating scales on which *E*s could rate their satisfaction with their participation in the experiment, their feelings about the animal *S*s, and their perception of their own behavior during the conduct of the experiment. Each scale ran from —10 (extremely dissatisfied) to +10 (extremely satisfied) with intermediate labeled points. On this questionnaire form, space was also provided for each *E* to describe how he felt before, during, and after the experiment.

Procedure

The experimental procedure is described in the instructions to *E*s. On the day the course instructor announced the details of the final experiment of the semester, the laboratory assistant entered the classroom announcing that the "Berkeley rats" had arrived. Instructions were read to *E*s and explained further where necessary. Each *E* was then asked to rate on a 20-point rating scale how much he or she expected to like working with the rats. (None of the *E*s had any prior experience with animal *S*s). On the basis of these ratings six pairs of *E*s were formed, matched on their liking of the rats. For each pair, one member was randomly assigned a group of *S*s which had been labeled "maze-bright" while the other member of the pair was assigned a group of *S*s which had been labeled "maze-dull." Actually, the groups had been labeled bright or dull randomly, with the restriction that differences in mean age per group per matched pair be at a minimum.

Before actually running any *S*s each *E* was asked to rate on a 20-point rating scale (+10, "extremely well," to —10, "extremely poorly") exactly how well he thought his *S*s would perform. Each of *E*'s five *S*s received one hour of handling and maze experience

before being run in the maze. During the maze experience, the *S* could obtain food from either arm of the T maze.

Each *E* ran each *S* ten times a day for five days and for each trial recorded whether the response was correct or incorrect and the time required to complete it. The darker arm of the maze was always reinforced while the white arm was never reinforced. The darker arm appeared equally often on the right and on the left, although the particular patterning of correct position was developed randomly for each day of the experiment and followed by all *E*s.

It was mentioned that twelve of the thirteen students in a particular course served as *E*s. The thirteenth student was an undergraduate research assistant who had worked for almost a year on the program of research on experimenter bias. While it seemed unlikely that any of the students in the class knew about the existence of this research project and the thirteenth student's connection with it, steps were taken to minimize the likelihood that such a connection could be made. The undergraduate research assistant therefore participated in the experiment along with the *E*s, but with the fully conscious motivation to get as good performance from her animal *S*s as possible without cheating. An advantage of her presence in this class was that since the course instructor rarely observed the actual conduct of the course experiments, she could serve as an informant on experimental procedures actually employed by *E*s without arousing the suspicion that might have been incurred had the instructor observed the experimental procedures. After the end of the semester during which this experiment took place, one of the *E*s was also invited to work with the research program on *E* bias. This *E* was thus also able to give valuable information on actual procedures employed by the *E*s during the conduct of the experiment. All reports made by these assistants to the senior author were held in confidence and no names of specific *E*s were mentioned.

RESULTS

Table 1 shows the mean number of correct responses per *S* for the six *E*s who believed they were running maze-bright *S*s, the six *E*s who believed they were running maze-dull *S*s, and the research assistant *E* who was aware that the *S*s were neither bright nor dull but who was trying to obtain good performance from her *S*s. Performance of the *S*s run by *E*s believing *S*s to be bright was significantly better on days 1, 4, and 5 but not on days 2 and 3. In addition, when the data from all five days of the experiment were combined, t was again significant, this time with a one-tailed p of .01.

TABLE 1
NUMBER OF CORRECT RESPONSES PER *S* PER DAY

N of *E*s	1	6	6		
N of *S*s	5	30	30		
Day	Asst.	Bright	Dull	t	p (one-tailed)
1	1.20	1.33	0.73	2.54	.03
2	3.00	1.60	1.10	1.02	NS
3	3.80	2.60	2.23	0.29	NS
4	3.40	2.83	1.83	2.28	.05
5	3.60	3.26	1.83	2.37	.03
Mean	3.00	2.32	1.54	4.01	.01

Inspection of the day-by-day means for each group of *E*s reveals that the "bright" *S*s' performance showed a monotonically increasing function such as might be expected if learning were occurring. The obtained monotonic increase could be expected by chance only six times in a hundred. The "dull" *S*s' performance, on the other hand, increased only to day 3, dropping on the fourth day and not changing on the fifth. The differences in obtained functions and the differences between performance means suggest that learning was less likely among *S*s run by *E*s believing them to be dull.

Table 1 also shows that except for the first day of the experiment, the *E* who was a research assistant and trying explicitly to obtain good performance from her *S*s actually did obtain better performance than did the *E*s believing their *S*s to be bright. The t for all five days was 2.38, which with four df was significant at the .05 level, one-tailed test. While her obtained performance function was not a monotonically increasing one, interpretation of this seems restricted by the relatively fewer *S*s run by this *E* compared to the two experimental groups. Interpretation of the

obtained t suggests that an E who is explicitly biased to obtain good performance from animal Ss obtains better performance than do Es who are biased to expect good performance but not explicitly instructed to obtain it.

TABLE 2
MEAN TIME IN MINUTES REQUIRED TO MAKE CORRECT RESPONSES

Day	Asst.	Bright	Dull	t	p (one-tailed)
1	5.45	3.13	3.99	NS	
2	1.63	2.75	4.76	NS	
3	2.04	2.05	3.20	NS	
4	0.74	2.09	2.18	NS	
5	0.68	1.75	3.20	NS	
Mean	2.11	2.35	3.47	3.50	.02

Of the 300 occasions when an S was run (60 Ss $\times$ 5 days), there were 60 occasions when the animal made no response at all. On the average, then, in one out of every five sessions the S refused to make a choice. This relatively poor performance may have been due to the difficulty of the discrimination problem, the limiting of pretraining to one hour, or the inexperience of the Es in running animal Ss. At any rate, these no-response occasions were not equally distributed between the experimental groups. There were 17 such occasions among the "bright" Ss and 43 among the "dull" Ss, a division which was significant at the .001 level (chi-square = 11.27). Since the "dull" Ss made fewer responses, and since Ss responding more are likely to respond correctly more often, it is possible that the results given in Table 1 were confounded. This likelihood of confounding would not, of course, account for the monotonic-nonmonotonic difference in the performance functions. In order to partial out the effects of greater nonresponding among the "dull" Ss, the mean time in minutes required to make only correct responses was computed for each day separately for the two experimental groups. The obtained mean times are shown in Table 2. Although for any given day the running times do not differ significantly between the two treatment groups, the difference for the entire experiment was found to be significant. Thus, Ss run by Es believing them to be bright make their correct choices more rapidly than do the Ss run by Es believing their Ss to be dull.

Inspection of the day-by-day means for the two treatment groups shows that the "bright" Ss show a more nearly monotonically decreasing function while the "dull" Ss perform more poorly on days 2 and 5 than on the just preceding days. The further question may be raised of whether "bright" Ss actually improved their performance compared to the "dull" Ss, or simply ran faster. Comparing the running time of the "dull" Ss on their first and fifth days yielded a t of less than one, suggesting that this group did not improve their performance significantly. The comparable t for the "bright" Ss was 1.77 which has a p of .06, one-tailed test, suggesting that this group probably did actually improve their performance during the course of the experiment.

Table 2 also shows that the E who was actually a research assistant obtained the shortest mean running time per correct response. Except for day 1, on which her Ss ran slowest of any group, her Ss performed better than did the Ss run by Es believing their Ss to be bright. This trend serves to support the earlier interpretation that an E who is explicitly biased to obtain good performance from animal Ss obtains better performance than do less explicitly biased Es.

Could the obtained results have been due to actual cheating on the part of the Es? The two Es who subsequently worked with the senior author on the research program on experimenter bias had been in a position to observe most of the Es' actual experimental procedures. There were no instances of rats not being run or of false entries on the data sheets. There were, however, a total of five observed instances of cheating in which an E prodded an S to run the maze. Two of these instances occurred among Es running "bright" Ss, while three occurred among Es running "dull" Ss. It appears unlikely from this distribution of instances of cheating that the differences obtained between the treatment groups could be ascribed to actual cheating behavior on the part of Es.

A question of some interest is whether both

groups of *E*s were successfully biased or whether only one of the groups was actually biased, with the other group obtaining data no different from what they might have obtained had they been unbiased. In several earlier studies in experimenter bias, in which human *S*s had been employed, the magnitude of the correlation between the data *E*s expected to obtain and the data they actually did obtain was used as an index of degree of experimenter bias (Rosenthal, Fode, & Vikan, 1960; Rosenthal, Friedman, Johnson, Fode, Schill, White, & Vikan, 1960). Since prior to running their *S*s all *E*s in the present study had been asked to predict the actual performance they expected to obtain from their *S*s, it was possible to employ this index of degree of bias. The Spearman rank correlation between expected and obtained performance was .43 for the *E*s running "bright" *S*s and .41 for those running "dull" *S*s. Since there were only six *E*s in each group, these correlations did not reach statistical significance, although when the groups were combined the one-tailed p reached the .09 level. These findings suggest that the two groups of *E*s were probably biased to about the same degree, although of course in opposite directions.

Table 3 presents the comparisons between the two treatment groups' ratings on the 23 scales of the postexperimental questionnaire. It will be noted that 14 of the 17 scales describing *E*s' perceptions of their own behavior during the conduct of the experiment have been grouped together into three clusters. These clusters were obtained and found to be statistically significant in an earlier study (Rosenthal, Fode, Friedman, & Vikan, 1960).

However, not all of the scales originally comprising the clusters were included in the questionnaire used in this study. The five scales dealing with *E*s' perceptions of *S*s were also grouped together to permit more meaningful summarization and discussion.

As described earlier, each scale was bipolar and is listed in Table 3 by the more desirable-sounding polar adjective. It should be mentioned, however, that some of these more desirable-sounding adjectives appeared over negative numbers in order to reduce the problem of response set. The most striking finding was that in 21 of the 23 scales, the *E*s believing their *S*s to be brighter rated them and their own behavior during the experiment more favorably. These *E*s tended to be more satisfied with their participation in the experiment, and to see their *S*s, on the whole, as brighter, cleaner, tamer, and more pleasant. These same *E*s rating themselves on the variables in Cluster 1 tended to describe their own behavior as more relaxed, casual, pleasant-voiced, and pleasant. These variables seem to describe a configuration which was earlier dubbed the "Perry Como Cluster" (Rosenthal, Fode, Friedman, & Vikan, 1960). On the variables in Clusters 2 and 3, *E*s running "bright" *S*s tended to see themselves

TABLE 3
MEAN RATINGS OF *S*s AND SELF

	Bright	Dull	t	$p < .20$ (two-tailed)
Satisfaction with experiment	3.0	2.5	2.10	.10
Ratings of *S*s				
Bright	4.2	—3.0	2.94	.04
Clean	7.2	2.2	1.24	
Tame	6.8	4.8	1.89	.13
Pleasant	4.8	0.0	1.77	.15
Like	4.8	2.2	0.92	
Mean	5.6	1.2	4.62	.01
Self Ratings				
Cluster 1				
Honest	3.8	3.7	0.16	
Relaxed	8.7	4.8	5.11	.006
Casual	6.8	3.3	1.33	
Business-like	3.7	5.3	—0.59	
Pleasant-voiced	7.3	3.7	1.65	.17
Behaved consistently	5.2	3.2	0.57	
Pleasant	6.7	2.8	2.56	.05
Mean	6.0	3.8	2.68	.04
Cluster 2				
Friendly	5.3	1.3	2.61	.05
Interested	6.8	6.3	0.54	
Encouraging	6.3	1.7	1.66	.17
Personal	0.7	2.2	—0.44	
Mean	4.8	2.9	1.31	
Cluster 3				
Nontalkative	6.2	3.2	1.19	
Enthusiastic	5.5	0.2	1.51	.19
Professional	2.5	—1.7	1.69	.16
Mean	4.7	0.6	6.21	.03
Nonloud	4.5	3.3	0.37	
Gentle handling of *S*s	6.5	2.7	1.95	.11
Much handling of *S*s	5.2	0.3	1.17	

as more friendly, encouraging, less talkative, more enthusiastic, and more professional. Finally, these *E*s saw themselves as handling their rats more, and more gently, than did the *E*s running "dull" rats. The mean difference between these two scales combined was significant at the .09 level, two-tailed test ($t = 7.91$, $df = 1$). If the accuracy of *E* self-ratings is likely, and an earlier study suggests that it may well be (Rosenthal, Fode, Friedman, & Vikan, 1960), it appears that these obtained differences in handling patterns may play a role in the mediation of experimenter bias to animal *S*s.

The solicited but unstructured comments made by all *E*s at the end of their questionnaire revealed that nine of the twelve *E*s felt good about *S*s' performing well and/or badly about *S*s' performing poorly. This appeared about equally true in both groups of *E*s, with four such comments occurring in the "dull" group and five in the "bright" group. Since even the *E*s in the "dull" group stated that they felt better if *S*s performed better, it appears that the mechanisms mediating the experimenter bias might not have been operating at a level of awareness available to *E*s. In fact, our obtained differences in performance seem to be more striking when it appears that the *E*s of both groups were very likely eager to obtain good performances from their *S*s.

DISCUSSION

That *E*s may bias their animal *S*s has been often discussed, and several workers have even referred, perhaps not entirely facetiously, to the *E*'s PK ability (Ammons & Ammons, 1957; Rotter[2]). This study suggests that not only does this biasing of animal *S*s occur now and then, but that it can be systematically induced and demonstrated.

That differences in animal handling were found to be related to the *S*s' performance should surprise no one. On a very gross level it might be hypothesized that researchers observing the manner in which a colleague removes a rat from a maze could judge significantly better than chance whether or not that *S* had performed as *E* had hoped. On a more subtle level, perhaps an *E*'s best judge of whether *E* is satisfied with *S* is *S*. An extra pat or two for a good performance, a none-too-gentle toss into the home cage for poor performance (where good and poor performances are defined by *E*'s hypothesis), may be very revealing to *S*. But, it may be said, no "good" researcher would do these things; a point which we may grant. While we know little of more subtle cues to animal *S*s, it does not seem farfetched to hypothesize that any *E* may react differentially to a well or poorly performing *S*; and this reaction, mediated by the autonomic nervous system, could well be transmitted to the animal *S* via changes in skin moisture, temperature, and the like. Thus grossly or subtly, Pavlov may well have been correct when he spoke of an *E*'s unwitting education of his mouse *S*s. This particular *E*, an assistant of Pavlov's, apparently ended his scientific career at this point of "error" (Razran, 1959), an event most reminiscent of Kinnebrook's dismissal by Maskelyne for the former's failure to observe stellar transits as quickly as the latter.

Whether many of the discrepant findings emerging from different laboratories with different and often opposite hypotheses may be accounted for on the basis of the experimenter bias phenomenon remains a moot point. One implication for research methodology does emerge, however. Whenever possible, the actual running of *S*s should be done by research assistants who do not know what outcome is desired. While this may be inconvenient to the researcher, it seems essential to rule out the operation of experimenter bias. Although we do not tell our assistant the hypothesis (we do not thereby deprive him of thought), he may try to guess the hypothesis. By chance he may bias *S*s in the expected direction half of the time, but our safeguard is that he will also bias them in the wrong direction half of the time. This, of course, increases the likelihood of Type II errors. However, it is suggested that to run our *S*s knowing the hypothesis is to increase the likelihood of Type I errors.

One of our reported findings must be qualified here. Our research assistant trying explicitly to influence her *S*s succeeded in

[2] Personal communication, 1960.

influencing them more than did the less explicitly biased *E*s. Findings from two earlier studies suggest, however, that *E*s who are explicitly biased or offered higher rewards for successful biasing may interpret this as a bribe by the researcher and actually bend over backwards to avoid biasing their *S*s (Rosenthal, Fode, & Vikan, 1960; Rosenthal, Friedman, Johnson, Fode, Schill, White, & Vikan, 1960).

This study is only a beginning in the research needed in the area of experimenter bias effects upon animal *S*s. We need to learn much about the effects on experimenter bias of various *E* and *S* characteristics. We need also to study further the mode of mediation of this bias including the role of auditory, visual, and tactile cues to the animal.

REFERENCES

Ammons, R. B., & Ammons, C. H. ESP and PK: A way out? *North Dakota Quart.*, 1957, 25, 119-121.

Ehrenfreund, D. A study of the transposition gradient. *J. exp. Psychol.*, 1952, 43, 81-87.

Fattu, N. A., & Mech, E. V. Interruption: its effect upon performance in a "troubleshooting" situation. *J. Psychol*, 1953, 36, 153-163.

Fode, K. L. The effect of non-visual and non-verbal interaction on experimenter bias. Unpublished master's thesis, University of North Dakota, 1960.

Gruenberg, B. C. *The story of evolution.* New York: Van Nostrand, 1929.

Maier, N. R. F. Frustration theory: restatement and extension. *Psychol. Rev.*, 1956, 63, 370-388.

Razran, G. Pavlov the empiricist. *Science*, 1959, 130, 916.

Rosenthal, R., & Fode, K. L. The problem of experimenter outcome-bias. In D. P. Ray (Ed.), *Series research in social psychology.* Washington, D. C.: National Institute of Social and Behavioral Science, 1961.

Rosenthal, R., Fode, K. L., Friedman, C. J., & Vikan, Linda L. Subjects' perception of their experimenter under conditions of experimenter bias. *Percept. mot. Skills*, 1960, 11, 325-331.

Rosenthal, R., Fode, K. L., & Vikan, Linda L. The effect on experimenter bias of varying levels of motivation of *E*s and *S*s. Unpublished manuscript, University of North Dakota, 1960.

Rosenthal, R., Friedman, C. J., Johnson, C., Fode, K. L., Schill, T. R., White, R., & Vikan, Linda L. Variables affecting experimenter bias in a group situation. Unpublished manuscript, University of North Dakota, 1960.

Zirkle, C. Pavlov's beliefs. *Science*, 1958, 128, 1476.

(Manuscript received October 22, 1962)

I said that a scientist is a man who goes and looks. Now I can explain what I meant. A scientist goes and looks, but his searching is not limited to material objects. He searches through ideas as well as through objects in order to find what he seeks. And he does not look indiscriminately—always he carries an image of what he seeks. What is he looking for? He is looking for what we all learned to look for in the first year of life. He is looking for something that matches up to his image of what the world must be, something that meets a test he himself imposes, something that has meaning only in terms of the standards he lives by. In that sense, the scientist is Everyman, looking just as you and I. We go and look for the things we want, and when we find them we find part of ourselves.

GEORGE A. MILLER, *Thinking, Cognition, and Learning*