Perceptual and Motor Skills, 1965, 20, 1237-1252. © Southern Universities Press 1965
Monograph Supplement 6-V20

PSYCHOLOGY OF THE SCIENTIST: XIV. EXPERIMENTERS' HYPOTHESIS-CONFIRMATION AND MOOD AS DETERMINANTS OF EXPERIMENTAL RESULTS[1]

ROBERT ROSENTHAL, PAUL KOHN, PATRICIA M. GREENFIELD, AND NOEL CAROTA

Harvard University

CONTENTS

Summary.—26 *Es*, each running about 6 *Ss* on a photo-rating task, were led to expect one of two opposite experimental results. Within each of these two groups of *Es*, half had their expectancies confirmed and half had their expectancies disconfirmed by their first 2 *Ss* (who were actually accomplices). Within each condition, half of the *Es* were praised and half were reproved for their experimental technique by one of two critics. The data obtained by any given *E* were found to depend on: (1) which of two critics had supervised him; (2) whether he was praised or reproved; (3) whether his early data returns confirmed or disconfirmed his initial hypothesis; (4) the initial hypothesis itself (when confirmed by early returns); and (5) certain of his more enduring personal characteristics.

A recently reported study showed that psychological data obtained by an *E* early in his experiment significantly influenced the data he obtained later in his experiment (Rosenthal, Persinger, Vikan-Kline, & Fode, 1963). In that study, three groups of *Es* were all led to expect similar results on a photo-rating task, but received "early returns" which varied in degree of correspondence to this expectancy. *Es* in the first group were given "good" data by two accomplices of the authors; *Es* in the second group were given "bad" data by two accomplices; and *Es* in a control group collected their initial data from "real" *Ss*. Upon running subsequent "real" *Ss*, this last group obtained data intermediate to those collected by *Es* in the two experimental conditions. Their results were neither so "good" (i.e., in accordance with their expectations) as those obtained by *Es* in the good early returns condition nor so "bad" as those collected by *Es* in the bad early returns condition.

The major purpose of the present study was to evaluate the contribution of two factors, expectancy and mood, to the mediation of the early returns effect. In the experiment summarized it is clear that initial confirmation of an expect-

[1]This investigation was supported by research Grants G-24826 and GS-177 from the Division of Social Sciences of the National Science Foundation. We want to thank Dr. Rudolf Kalin for his helpful reading of this paper and Dr. Winifred Lair for her invaluable assistance in obtaining *Ss*.

ancy by early data might have increased the strength of the expectancy and led to data more in accord with *E*'s expectancy or hypothesis. At the same time, however, confirmation might have improved *E*'s mood (Ebbinghaus, 1913; Griffith, 1961; Carlsmith & Aronson, 1963). If *E*'s mood were the more important mediating variable, the way in which it would have operated is not so clear. Since the experimental task in the earlier study was to rate pictures of human faces for the amount of success or failure the people pictured had been experiencing, a good mood could have led to more success ratings as such rather than to ratings which confirmed the hypothesis. The expectation induced in *E*s was that they would obtain *success* ratings from their *S*s. Thus, *E*'s mood might have operated in either of two ways: (1) by increasing his tendency to obtain confirmatory data in general and (2) by increasing his tendency to obtain success ratings whether or not these confirmed his hypothesis. Therefore, a third variable investigated in the present study was the nature of the initial expectancy induced in *E*. Thus, half of *E*s were led to expect success ratings from their *S*s, while the other half were led to expect failure ratings.

The confirmation or disconfirmation of expectancy variable in the present study was manipulated by the use of accomplices serving as the first two *S*s. These accomplices gave data either concordant or discordant with *E*'s initial expectancy or hypothesis.

Mood or hedonic tone was experimentally manipulated by having one of the authors praise or reprove each *E* for his technique of running *S*s *after* the accomplices had performed the experimental task but *before* the "real" *S*s had been put through the procedure. Praise was intended to induce a good mood in *E*, reproof a bad mood. Because we felt that the effects of such criticism might depend upon the personality of the critic and because we wanted to assess the generality of any obtained effects of criticism, we employed two critics rather than one (PK and RR).

An additional purpose of this study was to test further the general hypothesis that various *E* and *S* attributes might serve as unintended partial determinants of the results of psychological research (Rosenthal, Persinger, Vikan-Kline, & Mulry, 1963). More specific questions included: (1) How do *E*s' personalities influence the photo-ratings they obtain from their *S*s? (2) What kind of *E*s are particularly likely unintentionally to bias their *S*s' performance in the direction of their hypotheses? (3) What kinds of *S*s are particularly susceptible to such unintentional biasing effects? (4) How do the experiences an *E* has recently undergone (e.g., praise vs reproof, good vs bad early returns) affect his *S*s' personality test scores and personality test reliabilities?

METHOD

Samples

Experimenters.—Twenty-six Harvard College seniors served as *E*s. Twenty-five of them were majoring in Social Relations and writing honors theses. The twenty-sixth was a History major. All were volunteers from a target population of about 45 honors seniors in Social Relations.

Subjects.—*S*s were 115 female undergraduate students at a nearby college of elementary education. All were volunteers, although their psychology instructor encouraged students to participate. The sample was drawn from a total population of about 450 students.

Accomplices.—The accomplices who played the role of *S*s were students at another women's college and were selected on the basis of "trustworthiness." That is, they were well known to one of the authors before the experiment began. Twelve accomplices in all participated in the experiment, which was run on two evenings. Eight were used each evening, and six participated both evenings. Each accomplice served as "*S*" for 3 or 4 *E*s an evening.

Experimental Task

Each *E* presented each of his *S*s individually with a standardized series of 10 photos of human faces (Rosenthal & Fode, 1963). He asked his *S* to rate the photos for the degree of success or failure that she judged the person pictured to have been experiencing. *S*s used a standard rating scale which went from −10 (extreme failure) to +10 (extreme success) with intermediate labeled points.

Procedure

Four groups of *E*s ran *S*s at two different times on two evenings, two days apart. Before running their *S*s, each group of *E*s was given instructions by one of the authors who did not know to what condition any *E* would be assigned. Instructions included a demonstration of the experimental procedure. The experiment was presented to *E*s as a study of the personality correlates of research ability in the social sciences. After this training, *E*s were sent to their separate experimental rooms where they were to read their instructions, take some personality tests, and run their *S*s.

The personality tests included a birth-order questionnaire, the Taylor Manifest Anxiety Scale (MAS), the Marlowe-Crowne Social Desirability Scale (MCSD), and a TAT designed to assess *n* power, *n* achievement, and *n* affiliation.[2] The MAS and MCSD items were combined into one questionnaire which was split in half. The TAT pictures were also divided into two sets. *E*s did half of the MAS, MCSD, and TAT before running *S*s and half afterwards. The order of the two half-sets of tests was counterbalanced.

The written instructions to *E*s and the instructions they were to read to their *S*s were placed in the experimental rooms. Instructions were very similar

[2]The TAT was administered in connection with another study by other workers and is mentioned here only to give a fuller picture of the procedures used.

to those employed in the previous study of early data returns. The instructions to *E*s told how to administer the photo-rating task and cautioned against deviating from the exact instructions to *S*s. Also contained in the instructions to *E*s were the means for inducing the two initial expectancies. Half the *E*s were told that the particular type of *S*s they would be running had given average photo-ratings of about +5 in earlier research. The other half of the *E*s, on the other hand, were led to expect ratings of −5. Actually, of course, all *S*s were assigned to their *E*s in random fashion.

The first two "*S*s" run by each *E* were accomplices. Each accomplice was instructed by one of the authors to give photo-ratings averaging as close to +5 or −5 as possible (without using the same numbers suspiciously often). In half the conditions these ratings confirmed the expectancy previously induced in *E* ("good" early returns). In the other half, the accomplices' ratings disconfirmed the initial expectancy ("bad" early returns).

After *E*s had run their first two "*S*s" (the accomplices), one of the two critic-authors entered each experimental room and *praised* or *reproved E*. This manipulation was used to induce a good or pleasant mood in half the *E*s, and a bad or unpleasant mood in the other half. The critics were unaware of *E*s' initial hypothesis and the nature of their early data returns.

In the praise conditions, the critic entered, picked up *E*'s data sheets, studied them first with wrinkled brow and then with an increasingly pleased expression, and smiling at *E*, finally said approximately the following:

Your data follow an almost classical pattern. Haven't seen results that good in a long time. I'd tell you more specifically what's so good about them, except that it wouldn't be really cricket to do that now—perhaps later. Anyway, I'm sure you must be running things very competently to draw data patterns like that. Obviously, you've run *S*s before this. Well, keep up the good work with the rest of them. See you later.

In the criticism conditions, the critic entered, picked up *E*'s data sheets, studied them with wrinkled brow for about 30 sec., began to frown, and then said approximately the following:

Your data certainly follow a strange pattern. Haven't seen results like *those* in a long time. I'd tell you more specifically what bothers me except that it wouldn't be really cricket to do that now—perhaps later. Anyway, I'm sure you must be doing something strange to draw data patterns like that! I don't imagine you've run *S*s before this. Maybe empirical research is not your cup of tea. Well, please try to be very careful for the rest of them. See you later.

*E*s then ran from 3 to 6 "real" *S*s in succession. After that, they completed the personality tests. Finally, those *E*s who had been reproved were told that they really had done a very good job.

*S*s were given the same personality tests, except for the TAT; they also completed half the testing before and half after their photo-rating task. The order of the two half-sets of tests was counterbalanced, as for *E*s.

Experimental Design

Combination of the four variables described earlier—(1) +5 or −5 initial expectancy, (2) confirmation or disconfirmation of expectancy (i.e., "good" or "bad" early returns), (3) praise or reproof of *E*, and (4) Critic 1 or Critic 2—yielded 16 experimental conditions arranged in a 2 × 2 × 2 × 2 factorial design.

Precautions against authors' expectancy effects.—There was no way in which we could be certain that our own expectancies might not affect the results of this study. The following steps were taken, however, in the hope of minimizing any expectancy effects we ourselves might have on the data. (1) The author who instructed groups of *E*s did not know to which condition any *E* would be assigned. (2) Research rooms were randomly assigned to conditions. (3) *E*s were randomly assigned to research rooms. (4) Accomplices were randomly assigned to conditions (except that no accomplice could serve as *S* for an *E* known to her). (5) Accomplices did not know what the treatment conditions of the experiment were. (6) Critics were randomly assigned to conditions (except that number of praises and reproofs which each administered was equalized as closely as possible). (7) Critics did not know the particular condition of initial expectancy or of confirmation-disconfirmation for any *E* they contacted.

RESULTS

Effects of Single Conditions

The analysis of variance of the photo-ratings given by real *S*s is shown in

TABLE 1
ANALYSIS OF VARIANCE OF PHOTO-RATING DATA

Source	*df*	*MS*	*F*>2	*P*
Initial Expectancy (A)	1	0.183		
Confirm.-Disconfirmation (B)	1	2.288	3.95	.05
Praise-Reproof (C)	1	2.881	4.98	.04
Critic 1-2 (D)	1	2.488	4.30	.05
A B	1	1.271	2.20	.16
A C	1	0.158		
A D	1	0.001		
B C	1	0.432		
B D	1	1.789	3.09	.09
C D	1	0.002		
A B C	1	0.640		
A B D	1	1.375	2.38	.14
A C D	1	1.692	2.92	.10
B C D	1	0.316		
A B C D	1	0.116		
Error	99	0.579		

Table 1.[3] Three of the four main effects reached the .05 level. First of all, those *Es* whose expectations had been confirmed obtained *lower* ratings from their *Ss* than did *Es* whose expectations had been disconfirmed. Secondly, *Es* who had been praised by their critics obtained *lower* ratings from their *Ss* than did *Es* who had been reproved. Earlier we suggested that an improved mood might lead an *E* to obtain photo-ratings of greater success whether or not these confirmed his hypothesis. This did not turn out to be the case. Finally, *Es* who had been contacted by Critic 1 obtained lower ratings from their *Ss* than did *Es* contacted by Critic 2. This held true regardless of whether the critic was praising or reproving. Too much should not be made of these three main effects for two reasons: (1) they are not directly relevant to our major question; and (2) they are not independent of interaction effects which, while they did not reach the .05 level, cannot be safely disregarded both because of their magnitude and because they are the predicted effects the analysis of which is our major interest.[4]

Hypothesis confirmation.—Table 2 shows the mean photo-ratings obtained by *Es* in each condition of initial expectancy (+5 and −5) under conditions of

TABLE 2
MEAN PHOTO-RATINGS OBTAINED UNDER EACH INITIAL EXPECTANCY AND LEVEL OF CONFIRMATION

Level	Initial Expectancy	
	+5	−5
Hypothesis Confirmation	−1.16	−1.94
Hypothesis Disconfirmation	−0.96	−0.62

expectancy confirmation and expectancy disconfirmation (good and bad early data returns). Magnitude of experimenter bias, defined as the difference between data obtained by *Es* expecting +5 and those expecting −5 data, was significantly different for the two early returns conditions. When early returns were good, *Es* expecting higher ratings obtained higher ratings ($t = 1.71$, $p = .05$, one-tailed). However, when early returns disconfirmed *Es*' initial expectancies, they tended to obtain data discordant with their initial expectations but concordant with the data they obtained from *Ss* run early (accomplices). This tendency did not approach statistical significance, however. Among *Es* expecting +5 data, there was no significant difference in the data obtained by those

[3]Because of nonproportionality of *Ns* per semi-cubicle, the procedure suggested by Walker and Lev (1953) was followed. The effect of this procedure is to reduce *Fs* relative to their magnitude if *Ns* per cell were equal. We want to thank Ray Mulry for his discussion of this problem with us.

[4]The interactions of greatest interest are those involving A and B. Of these interactions the quadruple was not interpreted because it met neither of the two alternative required standards for interpretation: (1) either an *F* of reasonable magnitude or (2) planned within-interaction comparisons (Snedecor, 1956).

whose expectations were confirmed and that obtained by *Es* whose expectations were disconfirmed, although there was a tendency for *Es* to obtain higher ratings under the disconfirmation condition. This reverses the finding of the previous study (Rosenthal, Persinger, Vikan-Kline, & Fode, 1963) that *Es* expecting +5 data obtained significantly higher data when their early returns were good than when they were bad. The unexpected reversal of the earlier finding in the present study was attributable to the significant main effect of expectancy confirmation-disconfirmation.

Among *Es* expecting −5 data, lower data were obtained when expectancies were confirmed than when they were disconfirmed ($t = 2.66$, $p < .01$, two-tailed). From this analysis, we may conclude that *E*'s initial expectancy in part determined the data he obtained from his *Ss* when that expectancy was confirmed by early data returns. We may also speculate that, when the initial expectancy is disconfirmed by contrasting early data returns, these returns tend to replace the initial hypothesis as the expectancy salient to the running of subsequent *Ss*.

In this section we have shown the effects of hypothesis-confirmation without consideration of the effects of either *E*'s mood or critic differences. Table 1 shows that *E*'s mood did not interact with the effects of hypothesis confirmation. However, differences between our two critics did tend to complicate the effects of hypothesis confirmation. Table 3 shows the effects of hypothesis confirmation under conditions of contact with Critics 1 vs 2. For Critic 1 alone, *Es* whose expectancies were confirmed obtained significantly higher ratings when their initial expectancy was +5 rather than −5 ($t = 2.02$, $p < .03$, one-tailed). For Critic 1 again, *Es* whose initial expectancies were disconfirmed, subsequently obtained data inconsistent with their initial expectation (but consistent with the data they obtained from their disconfirming *Ss* who were accomplices, $t = 1.36$, $p = .17$, two-tailed). For Critic 2 alone, *Es* initially expecting to obtain +5 data tended to obtain higher data than did those expecting −5 data regardless of whether their initial expectancy was confirmed or not. This tendency did not approach significance, however. From this we may conclude that individual differences among *Es*' critics may partially determine the nature of the

TABLE 3
MEAN PHOTO-RATINGS OBTAINED UNDER EACH INITIAL EXPECTANCY AND LEVEL OF CONFIRMATION AS A FUNCTION OF *E*'S CRITIC

Level	Initial Expectancy			
	Critic 1		Critic 2	
	+5	−5	+5	−5
Confirmation	−1.58	−2.96	−0.73	−0.90
Disconfirmation	−1.31	−0.39	−0.62	−0.84

effect of early data returns. Since both critics administered both praise and reproof, it follows that the content of the critics' message to *E* could not be the effective variable. Furthermore, since both critics were blind to *Es*' initial expectancy and to whether early data returns were confirmatory or disconfirmatory, the effect of critic differences cannot be attributed to an outcome-orientation bias on the critics' part (Rosenthal, 1964).

Mood effects.—We turn now to the effects on *E*'s unintentional biasing of his being praised or reproved for his role-behavior. These effects were found to depend on which of the critics had praised or reproved *E* ($p = .10$).

TABLE 4

MEAN PHOTO-RATINGS OBTAINED UNDER EACH INITIAL EXPECTANCY AND LEVEL OF CRITICISM AS A FUNCTION OF *E*'S CRITIC

E's Behavior	Initial Expectancy			
	Critic 1		Critic 2	
	+5	−5	+5	−5
Praise	−1.66	−2.34	−1.52	−0.86
Reproof	−1.24	−1.02	+0.17	−0.88

Table 4 shows mean photo-ratings obtained as a function of three variables: initial expectancy (+5 or −5), level of criticism (praise or reproof), and person assigned as critic. For Critic 1, *Es* obtained higher data when they expected higher data (ie., +5) only when they had been praised by the critic. On the other hand, when they had been reproved by him, they obtained slightly lower data when they expected +5 as compared to −5 data. Critic 1's *Es*, then, showed more initial expectancy-bias when praised than when reproved. The situation was reversed, however, with Critic 2. His *Es* initially expecting +5 data obtained significantly higher ratings than those expecting −5 data when they were reproved. On the other hand, when he praised them, his *Es* initially expecting +5 data obtained lower data than did those initially expecting −5 data. Critic 2's *Es*, then, showed the expected bias only when they were reproved but showed a reverse bias when praised. The effects of praise and reproof on unintentional biasing seem, then, to depend partly on who does the praising or reproving.

Critic effects.—Table 5 shows for each critic separately, the data his *Es* obtained when their expectancies were confirmed and disconfirmed. This analysis disregards initial expectancies as well as condition of praise or reproof. Once again, marked critic differences appear. For Critic 1, *Ss* run by *Es* whose expectancy had been confirmed rated the photos significantly lower than did *Ss* run by *Es* whose expectancy had been disconfirmed ($t = 3.13$, $p < .005$, two-tailed). This difference, while in the same direction for Critic 2, was not statistically significant.[5]

TABLE 5

MEAN PHOTO-RATINGS OBTAINED UNDER CONDITIONS OF CONFIRMATION AND DISCONFIRMATION AS A FUNCTION OF *E*'S CRITIC

Level	Critic 1	Critic 2
Confirmation	−2.28	−0.82
Disconfirmation	−0.85	−0.73

Early returns as sources of expectancy.—We noted earlier that, where early returns confirmed initial expectancies, *Es* showed the greatest biasing effect upon their *Ss*. Further, we found (at least for one of the two critics) that, when early returns disconfirmed the initial expectancy, *E*'s subsequent data would tend to be consistent with the early returns. We suggested that the early returns created an expectancy which supplanted the one initially induced. The plausibility of this explanation would be increased if we could show that early returns by themselves could induce an effective expectancy in *Es* which was somehow communicable to subsequent *Ss*. Here we present evidence to that effect. Within the eight cubicles remaining when the critic dimension was collapsed,[6] a product-moment correlation was computed between the magnitude of the mean ratings given to *Es* by their first two *Ss* (accomplices) and the magnitude of ratings subsequently obtained from real *Ss*. It should be noted that within any of these eight conditions the mean of the data given an *E* by our accomplices fluctuated only slightly (average deviation = 0.5) and more or less randomly from the +5 or −5 ratings accomplices had been instructed to approximate.

TABLE 6

CORRELATIONS BETWEEN DATA PRODUCED BY ACCOMPLICES AND DATA SUBSEQUENTLY PRODUCED BY REAL *Ss*

Level	Initial Expectancy			
	+5		−5	
	Praise	Reproof	Praise	Reproof
Confirmation	.67*	.69*	.73	.999
Disconfirmation	.88	−.76	.06	.44

*N of *Es* = 4; for all other cells N of *Es* = 3.

[5] In all preceding analyses, *df* were based on N of *Ss* per condition. Since 8 of the conditions employed more than a single *E*, 8 estimates of between-*E*, within-condition variance were available for testing against within-*E* (between-*S*) variance. All *Fs* clustered around unity ($M = 1.03$), 4 exceeding it and 4 not. We concluded that *Es* within conditions differ no more from each other in data obtained from randomly assigned *Ss* than *Ss* run by a single *E* differ among themselves. Mulry (1962) obtained similar results with a pursuit rotor.

[6] No differences were found between the correlations being reported; this could be a function of which critic had contacted *E*.

Table 6 shows the obtained correlations. Only one is not positive. The mean z transformed correlation was +.79 ($p = .002$, one-tailed, $df = 10$). Inspection of Table 6 suggests that this over-all correlation may mask a difference between those correlations obtained when accomplices were confirming as opposed to disconfirming *Es*' initial hypotheses or expectancies. When initial expectancies were being confirmed the mean r was +.96 ($p < .0005$, one-tailed, $df = 6$). However, when initial expectancies were being disconfirmed, the mean r (+.23) was not significantly greater than zero. It was, however, significantly lower than the mean r of +.96 ($z = 2.57$, $p = .01$, two-tailed). At least those *Es*, then, whose initial expectancies were confirmed by their early data returns tended to obtain data from *Ss* run later which were similar to the data obtained from *Ss* run earlier in spite of an artificially restricted range of early data returns. However, two quite different factors might have been operating: (1) an *E*'s personality factor or (2) an *E*'s expectancy factor. If the personality factor were operant, *Es* would have affected the accomplices in the same way in which they subsequently affected their real *Ss*. Accomplices were, after all, free to vary in their ratings at least a little. If, on the other hand, the expectancy factor were operant, the data produced by the accomplices would serve to modify the original expectancy however minimally. The modified expectancy might substantially influence the data subsequently obtained from real *Ss*.

If the personality factor were operative, one would expect *Es* to have a relatively constant effect on the data they collected from accomplices and real *Ss* regardless of order; hence, the correlation between data obtained from accomplices and those from *Ss* run later should be no higher than the correlation between the data from *Ss* run early and those run late. The correlation between ratings given *Es* by accomplices and those given by real *Ss* subsequent to the first two was +.85 ($p = .005$, one-tailed, $df = 6$). The correlation between ratings given by the first two real *Ss* and those given by subsequently run real *Ss* was only .16, which does not differ appreciably from zero but is appreciably lower than +.85 ($p = .06$, two-tailed). While these findings seem inconsistent with the *E*'s personality hypothesis, they are not inconsistent with the *E*'s expectancy hypothesis. A reasonable conclusion might be that the earliest data returns, if consistent with *E*'s initial expectancy, serve to specify expectancies further. Data given later in the collection process by the first two real *Ss* seem to have less effect on data obtained still later. This might be due in part to the greater psychological impact of data collected first in the experiment. In part too, it might be due to the real *Ss*' giving ratings averaging around zero and thereby possibly disconfirming *Es*' expectancies at least in contrast to the very confirming data given by the accomplices.

Delayed Action Effect

In the earlier experiment on the effect of early data returns we had found a "delayed action effect," with accomplices' data affecting real *Ss* run later more than those run earlier. In the present study no such effect was found. There was, however, a tendency for *Es*' initial expectancies to become more effective for later- than for earlier-run *Ss*. Initial expectancies (+5 vs −5) had no effect on data obtained from the first two real *Ss*. Among subsequently-run real *Ss*, however, those run by *Es* initially expecting +5 ratings gave higher ratings (−.35 compared to −1.21) than did *Ss* run by *Es* expecting −5 ratings ($t = 1.73$, $p = .09$, two-tailed, $df = 83$).

Accomplices as "Victims"

One final report on the behavior of our accomplices is in order. In social psychological research we are accustomed to think only of the influence our accomplices have on our "targets." That our targets influence our accomplices as well, while never denied, is too rarely asserted. Table 7 shows the mean absolute ratings given *Es* by the accomplices for each condition of initial expectancy and for each condition of expectancy confirmation-disconfirmation. Data given *Es* initially expecting −5 data and receiving confirming data from our accomplices, were significantly lower in absolute value than those given in the other three conditions ($t = 7.80$, $df = 2$, $p = .02$, two-tailed). In some way, *Es* in this condition did not "permit" our accomplices to come as close to their programmed responses as they had in the other conditions. It should be emphasized that the treatment conditions were in "*E*'s head" and not in the accomplices'. The latter were not only blind to their *Es*' treatment conditions, they did not even know the over-all purpose or design of the experiment.

TABLE 7
MEAN ABSOLUTE RATINGS (SIGNS DISREGARDED) GIVEN *Es* BY ACCOMPLICES UNDER FOUR TREATMENT CONDITIONS

Level	Initial Expectancy	
	+5	−5
Confirmation	4.07	3.54
Disconfirmation	4.16	4.20

Attributes of E *and* S *as Data Determinants*

E*'s attributes and* S*s' photo-ratings.*—Over all experimental conditions, first-born *Es* obtained significantly higher photo-ratings of success from their *Ss* than did later-born *Es* ($\chi^2 = 5.85$, $p = .02$, $df = 1$).

Again, over all experimental conditions, those *Es* who scored as less anxious on the combined pre- and post-MAS, obtained higher photo-ratings of success than did more anxious *Es* ($\chi^2 = 7.55$, $p < .01$, $df = 1$). This finding also emerged when the total MAS score was separated into the pre-test and post-test portions (ps = .13 and .01, respectively). Finally, *Es*' scores on the combined

pre- and post-MCSD and on the post- alone were not significantly related to the photo-ratings made by their *Ss*; however, *Es* earning higher MCSD scores on the pre-test obtained significantly higher photo-ratings of success than did those earning lower MCSD scores ($\chi^2 = 3.85$, $p = .05$, $df = 1$).

E's attributes and experimenter-bias.—To cross-validate earlier findings (Rosenthal, 1963) bearing on the relationship between *Es*' MAS and MCSD scores and their success in obtaining data concordant with their expectancies, we divided those 14 *Es* whose initial expectancies had been confirmed, into more effective and less effective "biasers." Effective "biasers" were those 7 *Es* who obtained mean data from their real *Ss* which agreed most clearly with their mean data obtained from accomplices within each experimental condition.[7]

Es scoring lower on the need for social approval variable (MCSD) biased their *Ss* more ($t = 2.38$, $p = .04$, two-tailed, $df = 12$). This was equally true regardless of *Es*' anxiety level. Medium-anxious *Es* tended to bias their *Ss* more than those earning very high or very low MAS scores ($t = 1.98$, $p = .08$, two-tailed, $df = 12$).

Subject susceptibility to experimenter-bias.—Among *Ss* run by *Es* whose early returns were confirmatory, i.e., the group among whom biasing effects clearly occurred, *S*'s susceptibility to *E*'s biasing effects was defined as follows. If *E*'s initial expectancy was for a +5, those of his *Ss* giving data higher than his mean obtained data were called more susceptible to biasing effects, while those of his *Ss* giving data lower than his mean obtained rating were called less susceptible. If *E*'s initial expectancy was for a −5, those of his *Ss* giving data lower than his mean obtained rating were called more susceptible to experimenter-bias effects, while those of his *Ss* giving data higher than his mean obtained data were considered less susceptible.

Among *Ss* run by both more and less biasing *Es*, there was no relationship between need for social approval (pre-test plus post-test MCSD) and susceptibility to experimenter-biasing effect.

Among *Ss* run by more biased *Es*, those who scored either high *or* low on anxiety (pre-test plus post-test MAS) were more susceptible to biasing effects ($t = 2.14$, $p = .05$, two-tailed, $df = 26$). Among *Ss* run by somewhat less biased *Es*, those who scored as less anxious were more susceptible to biasing effects ($t = 2.30$, $p = .14$, two-tailed, $df = 33$). For both the MAS and MCSD variables, the full scale scores were employed because there were no significant differences between pre- and post-test scores or in magnitude of shift scores as a function of degree of experimenter-biasing effects or subject-susceptibility to such biasing effects.

Experimenters' experiences and Ss' test-score changes.[8]—There were differences from pre- to post-test score changes as a function of which of our two critics had contacted *Ss*' *Es*. Thus, those *Ss* whose *Es* had been contacted by Critic 2 showed a significantly greater increase in MAS scores than did *Ss* whose *Es* had been contacted by Critic 1 ($\chi^2 = 7.71$, $p < .01$, $df = 1$). In addition, those *Ss* whose *Es*' initial expectancies had been confirmed showed a significant increase in their MCSD scores as compared with MCSD scores of those *Ss* whose *Es*' initial expectancies had been disconfirmed. This relationship, however, held only among *Ss* whose *Es* had been contacted by Critic 2 ($\chi^2 = 9.84$, $p < .01$, $df = 1$).

Finally, we calculated the test-retest reliabilities of *Ss*' MAS and MCSD scores separately for those run by (1) *Es* whose initial expectancy had been confirmed and (2) *Es* whose initial expectancy had been disconfirmed. The MAS reliability, corrected for length, was significantly higher among *Ss* whose *Es*' initial expectancies had been confirmed ($r = +.90$) rather than disconfirmed ($r = +.80$; z diff. $= 1.91$, $p = .06$, two-tailed). The corresponding MCSD reliabilities of .74 and .66 did not differ significantly from each other. None of the other experimental conditions affected the reliabilities of either the MAS or MCSD scales.

Discussion

Experimenter Experiences

The things that happen to an *E* during the conduct of his experiment seem clearly to affect the data he obtains from his *Ss*. Thus, if his initial hypothesis is confirmed by the early data returns, he will tend to obtain different data from subsequently-run *Ss* than he would if it is disconfirmed. If he is praised early in the data-collection process, he will obtain different subsequent data than if he is reproved. In addition, just who does this praising or reproving also affects the data *E* obtains from *Ss* run subsequently. These findings support the hypothesis that the data obtained by a psychological researcher are determined in part by what sort of person he currently is, including what has happened to him just before he contacts his *Ss*. Such factors affect the general order of magnitude of the data obtained from *Ss*. But insofar as such factors are not associated differentially with experimental conditions, the magnitude of the differences among experimental conditions should not be affected.

An *E*'s expectancy does, however, have significantly differential effects on data obtained under different experimental conditions. This was most clearly so when early data returns confirmed *E*'s hypothesis. This result serves as a general replication of our earlier finding on the effect of early data returns.[9]

[7] A more detailed discussion of the logic of several definitions of "bias" magnitude is presented elsewhere (Rosenthal, Persinger, Vikan-Kline, & Mulry, 1963).

[8] For both the MAS and MCSD measures, we analyzed the effect of our four experimental variables on the change from *Es*' pre- to post-test scores. None of the *F*s emerging from these $2 \times 2 \times 2 \times 2$ analyses of variance had an associated $p < .20$.

[9] The significant main effect of the confirmation variable in the present study, however, was unpredicted and at best interpretable only in *post hoc* terms.

Effects of Principal Investigators

The effect of early data returns was seen to operate primarily among *Es* who had been contacted by Critic 1. Extrapolating to the more usual research situation, our critics may be viewed as analogues of the principal investigator. It may be that the effects of early data returns operate only among data-collectors contacted by certain principal investigators. In addition, individual differences among principal investigators (critics) appear to interact with the nature of their contacts with *Es*. For some principal investigators, praise of *Es*' work tends to increase *Es*' expectancy-bias; for other principal investigators, reproof of *Es*' work tends to increase their biasing effect.

Psychological investigations employing a test-retest paradigm, including studies of test reliabilities, may be subject to some rather non-obvious sources of error. Individual differences among "principal investigators" differentially affected *Ss*' retest scores on the Taylor MAS. The results of studies employing a personality test-retest method, therefore, appear to be subject not only to the experiences and characteristics of data collectors but also to the personal characteristics of the principal investigators who supervise the data collectors. This effect of the principal investigator appeared to be mediated via *Es* under their supervision since the principal investigators themselves had essentially no direct contact with the *Ss*.

Accomplices as "Targets"

Our finding that *Es* were somehow able to affect the data given them by our accomplices raises some interesting questions. Perhaps when we employ accomplices in social psychological research, we overestimate the unidirectionality of influence from accomplice to "target." Apparently our "targets" are not simply at the mercy of our accomplices but have notable effects on them as well.

E*s' Attributes*

Certain personal attributes of *Es* may also serve as determinants of data obtained from *Ss*. In the present study, *Es* who were first-born, less anxious, and perhaps with higher need for social approval, obtained from their *Ss* significantly higher ratings of success of persons pictured in photos. None of these three relationships had been obtained under the analogous conditions of an earlier experiment (Rosenthal, Persinger, Vikan-Kline, & Mulry, 1963). In that study, in fact, the relationship between photo-ratings and *Es*' need for social approval (MCSD) tended to be in the opposite direction.

In five earlier studies of biasing effects, *Es* with a higher need for social approval were found to bias their *Ss* more (Rosenthal, 1963). Our finding just the opposite in the present study is difficult to explain but does not seem attributable to chance fluctuation.

The finding that medium-anxious *Es* bias more than either high- or low-anxious *Es* supports data of one earlier experiment but disagrees with those of two others. In one of these studies *high*-anxious *Es*, and in another study, *low*-anxious *Es*, were found to bias most (Rosenthal, 1963). Thus, both anxiety and need for social approval are related to magnitude of experimenter-biasing effects. Apparently, the nature of the relationship is dependent on some interacting but as yet unknown variable(s).

S*s' Attributes*

A similar statement may be made regarding *Ss*' attributes associated with greater or lesser susceptibility to experimenter-biasing effects. Of a total of six samples, there were two samples in which less anxious *Ss* were most susceptible to bias, two samples in which medium-anxious *Ss* were most susceptible, one sample in which highly anxious *Ss* were most susceptible, and one sample in which extreme (high or low) scorers were most susceptible (Rosenthal, 1963). At least one variable, *S*'s need for social approval, has consistently been found to be unrelated to biasability.

Conclusions and Implications

What this study makes quite clear is that *Es* may obtain significantly different data from their *Ss* on the basis of five extraneous but important factors: (1) who trains and supervises *E*; (2) what *E* is told about his skill as an *E*; (3) whether his early data are highly confirming or disconfirming of his initial hypothesis; (4) what he expects his data to be (provided that early data confirms his initial hypothesis); (5) certain of his more enduring personal characteristics.

If all of these can affect the data an *E* obtains, what are the implications for the gathering of "experimenter-effect-free" data in psychological research? The suggestions listed below are primarily designed to reduce the *systematic* biasing effects of *E*'s expectancies, but, in addition, should serve to reduce the more random but still serious error associated with individual differences among *Es*: (1) the elimination of *E-S* contact where the design of the research permits this; (2) the institution of effective double-blind procedures if *E-S* contact is required; (3) the employment of random samples of *Es* drawn from a relevant population of relevantly uninformed *Es* (This procedure would increase the generality of our data even if no experimenter-biasing effects were operant.); (4) the employment of systematically drawn *Es* with known or determinable distributions of expectancies; (5) the systematic purposeful induction of different expectancies in several samples of *Es*.

One specific implication for the interpretation of experimental results, especially in social psychological research, must be considered: that no experimental data in social psychology based on only one *E*'s contacting *Ss* should be taken at face value unless replicated by at least one different *E*. Any data obtained, including the differences among treatment conditions, may be as much a function of *E* as of the experimental manipulation.

REFERENCES

CARLSMITH, J. M., & ARONSON, E. Some hedonic consequences of the confirmation and disconfirmation of expectancies. *J. abnorm. soc. Psychol.*, 1963, 66, 151-156.

EBBINGHAUS, H. *Memory: a contribution to experimental psychology.* New York: Teachers College, Columbia Univer., 1913.

GRIFFITH, R. Rorschach water percepts: a study in conflicting results. *Amer. Psychologist*, 1961, 16, 307-311.

MULRY, R. C. The effects of the experimenter's perception of his own performance on subjects' performance in a pursuit rotor task. Unpublished master's thesis, Univer. of North Dakota, 1962.

ROSENTHAL, R. On the social psychology of the psychological experiment: the experimenter's hypothesis as unintended determinant of experimental results. *Amer. Scientist*, 1963, 51, 268-283.

ROSENTHAL, R. Experimenter outcome-orientation and the results of the psychological experiment. *Psychol. Bull.*, 1964, 61, 405-412.

ROSENTHAL, R., & FODE, K. L. Three experiments in experimenter bias. *Psychol. Rep.*, 1963, 12, 491-511.

ROSENTHAL, R., PERSINGER, G. W., VIKAN-KLINE, L., & FODE, K. L. The effect of early data returns on data subsequently obtained by outcome-biased experimenters. *Sociometry*, 1963, 26, 487-498.

ROSENTHAL, R., PERSINGER, G. W., VIKAN-KLINE, L., & MULRY, R. C. The role of the research assistant in the mediation of experimenter bias. *J. Pers.*, 1963, 31, 313-335.

SNEDECOR, G. W. *Statistical methods.* (5th ed.) Ames, Ia: Ia State College Press, 1956.

WALKER, H. M., & LEV, J. *Statistical inference.* New York: Holt, 1953.

Accepted April 19, 1965.