Journal of Educational Psychology
1969, Vol. 60, No. 5, 377-383

EXPECTANCY EFFECTS IN THE CLASSROOM: A FAILURE TO REPLICATE

WILLIAM L. CLAIBORN[1]
Syracuse University

Twelve first-grade classrooms were divided equally among four groups representing the combinations of presence or absence of the expectancy bias and classroom observers. In the bias classes, each teacher received a list of approximately 20% of her pupils who could be expected to show "intellectual blooming" when these pupils were in fact picked without regard to intellectual potential. Retesting 2 months later showed no relative gains for pupils who were the object of the expectancy bias; there were no clear changes in observed teacher-pupil interaction. Differences between the present study and previous studies were discussed in light of this "failure to replicate." It was concluded that the evidence for bias effects in the school remains equivocal.

Rosenthal and Jacobson (1968) and others have reported that changing a teacher's expectation of a particular pupil's intellectual potential results in changes in that pupil's performance on a standardized group IQ test. This research has grown out of a long line of *E* bias studies (Rosenthal, 1964, 1966, 1967) which have demonstrated that the expectations of *E* are a significant factor in determining the outcome of the experiment. This appears to occur with the greatest regularity where the demands of the task are relatively vague (Shames & Adair, 1967). However, Barber and Silver (1968) presented an analysis of 31 studies which have attempted to demonstrate the *E* bias phenomenon. According to their reanalysis and reinterpretation of the design and statistics presented in these studies, the majority did not clearly or unequivocally demonstrate the unintentional *E* bias effect. Barber and Silver are particularly critical of post hoc analyses following failure to reject the null hypothesis and of post hoc probability pyramiding. They conclude that the bias effect is more difficult to demonstrate and less pervasive than has been previously assumed.

A thorough reading and analysis of the recent book, *Pygmalion in the Classroom* (Rosenthal & Jacobson, 1968), shows that the same difficulties discussed by Barber and Silver (1968) are inherent in their presentation. Rosenthal and Jacobson randomly chose approximately 20% of the children in each of three classes in Grades 1–6. Teachers were told that these "special" children could be expected to show intellectual blooming in the coming months. What follows is an example of the kind of difficulty present in the authors' interpretation of their data. At the end of the first year, the special children had gained more IQ points relative to the control children ($p < .02$, one-tailed). This effect is largely attributable to substantial changes in one first-grade classroom in which the special children showed a relative advantage of 15.4 IQ points. There was no significant IQ gain reported for Grades 3–6. Thus, there was no teacher expectancy effect in two-thirds of the grades examined. More importantly, only 2 of the 18 classes (one first and one second grade) yielded any reliable IQ increase; one third grade showed a significant decrease. Examination of the pretest IQs for the special and nonspecial pupils in the three first grades shows differences in initial IQ. Significance tests indicated that one first grade differed on the verbal subtest; the special children had the lower pretest scores. When this −28 point initial discrepancy is compared with the post-

[1] Now at the University of Maryland. The article is based on the author's doctoral dissertation submitted to Syracuse University, while the author was a Public Health Service fellow. Requests for reprints should be sent to: William L. Claiborn, Department of Psychology, University of Maryland, College Park, Maryland 20742.

test discrepancy of +8 (*ns*) points, it is apparent that regression effects could account for the observed changes. It is this class which produces the significant IQ change score. Randomization failed to protect the selection of special children and resulted in uninterpretable effects. The authors' pre- to posttest difference-score analysis clearly does not permit unambiguous statements attributing changes to treatment. In a real sense, no expectancy effects can be claimed for the first grade.

Analyses of other results reported by Rosenthal and Jacobson (1968) make it reasonable to fail to reject the null hypothesis that no reliable teacher expectancy effects were observed. In any case, the findings upon which Rosenthal and Jacobson (1968) based their conclusions were difference scores, not corrected for known pretest differences, and partially attributable to regression effects. Further difficulties relating to post hoc hypothesis support, partial data analysis, and probability pyramiding are presented in Claiborn (1968).

Despite these weaknesses, Rosenthal and Jacobson (1968) conclude that telling a teacher that some of her children are likely to show intellectual blooming is sufficient to result in changes in the pupil's obtained IQ. It is implied that the measured changes are not artifacts of the experimental procedure.

> Teachers may have treated their children in a more pleasant, friendly, and encouraging fashion when they expected greater intellectual gains of them. Teachers probably watched their special children more closely, and this greater attention may have led to more rapid reinforcement of correct responses with a consequent increase in the pupil's learning.... Such communication together with possible changes in teaching technique may have helped the child learn by changing his self-concept, his expectations of his own behavior, and his motivation, as well as his cognitive style and skills [Rosenthal & Jacobson, 1968, p. 180].

The primary purpose of this present research was to observe and quantify some in-class teacher-pupil behavior in an attempt to "capture" changes in teacher behavior which would follow the introduction of a fictitious statement about the intellectual potential of some of her pupils. The second purpose of the study was to "replicate" the earlier purported finding that providing a teacher with a bias for the intellectual growth of some of her pupils results in the improvement in intellectual performance for those special pupils. Most generally it was hypothesized that altering the teacher's expectancy for the intellectual potential of some of her pupils would result in (*a*) an increase in the child's IQ as measured by difference between pre- and posttest on a standardized group IQ test; (*b*) a differential change in teaching behavior toward the children who were the objects of the expectancy bias. The biased teacher would have more frequent contact, express more positive affect, and would tend to expand upon the contributions of the special children.

METHOD

Design

The study can be described as a 2 × 2 factorial experiment. The two levels of the first factor consisted of the presence or absence of raters in the classroom; the two levels of the second factor consisted of the presence or absence of induced expectancies for intellectual blooming. As Table 1 illustrates, the research plan provided for four experimental groups. Each group consisted of three classrooms chosen from the available sample of 12 first grades.

In brief, in the beginning of the spring term a 2-week observation period in Rated classrooms was followed by a testing session in all classes. The results of the testing for "potential intellectual bloomers" was made known to the Bias classroom

TABLE 1

TREATMENT CONDITIONS FOR EACH OF THE FOUR EXPERIMENTAL GROUPS

Group	Classroom behavior ratings	Pretest	Expectancy bias	Classroom behavior ratings	Posttest
Rated—Bias	×	×	×	×	×
Rated—No Bias	×	×	0	×	×
Unrated—Bias	0	×	×	0	×
Unrated—No Bias	0	×	0	0	×

Note.—"×" indicates the presence of a treatment, "0" indicates its absence.

teachers following the first testing period; observation was continued for an additional period. Two months later, near the end of the school year, the IQ test was readministered to all children.

Sample

The schools from two predominately middle-class suburbs of a major upstate New York community each provided four first-grade classes. Each cell of the 2 × 2 (Rated—Bias; Rated—No Bias; Unrated—Bias; Unrated—No Bias) contained three classrooms, one from each of the three different schools. Within the school, classes were assigned to the Bias and No Bias conditions (with one exception) at random. Assignment of classes to Rated and Unrated conditions was made by the school principal. The mean pretest IQ for classes ranged from 92.4 to 118.4.

Within each classroom, approximately 20% (from four to five) pupils were designated as "special" and subject to special analysis. The special pupils in the Bias classrooms were presented to the teachers as "potential bloomers." The 20% were chosen proportionately from the males and females in the class, and within sex randomly, from the upper and lower half of the pretest IQ distributions.

Measures

The primary dependent measure was pre- to posttest IQ differences obtained from the group administration of Test of General Ability (TOGA) developed by Flanagan (1960). The limits at the high end of the tabled norms for this test necessitated extrapolations for IQ scores for a few pupils. A similar problem, though at the opposite end of IQ range, developed in the Rosenthal and Jacobson (1968) experiment.

The second major set of dependent variables was obtained from systematic ratings of classroom behavior.[2] Each classroom in the Rated condition was scheduled for regular observation sessions. The 20-minute sessions were divided into a pretest unit of 2 weeks and a posttest unit of 7 classroom days. The system of rating was designed to evaluate teacher interactions regarding particular pupils rather than to assess the teacher's general teaching behavior. Each interaction was scored for the particular pupil involved. The rating attempted to assess the nature and frequency of teacher-pupil interactions, including affective aspects.

Procedure

The period from the beginning of the classroom rating to the final posttesting was just under 3 months in the spring of a regular school year. In the beginning of the spring term, following some preliminary observation, the first formal rating period lasted 2 weeks and was introduced to the teachers as part of a requirement for a graduate education course. Following this observation period, children in each of the 12 classes were tested with the TOGA. Teachers were told that the test was designed to predict "intellectual blooming." At the end of the same week, "test results" were distributed to the teachers in the Bias classes only. These results in fact reported the names of the 20% who had been independently chosen from each class. Immediately following the testing and the introduction of the expectancy bias, classroom observation continued for an additional period of about 1½ weeks. In the first week of the last month of school, all classes were retested with the same IQ test. Teachers also completed a questionnaire which assessed their awareness of the nature of the experiment, and their ability to remember the names of the students who had been designated as "bloomers."

The classroom observers and the author were aware of the major experimental hypotheses. The teachers, however, were not told the nature of the experiment and the raters did not know which were the Bias and which were the No Bias classrooms. In addition, only the Biased teachers and the author were aware of which children were designated as "potential bloomers."

RESULTS

From the questionnaires it was evident that the teachers were able to accurately remember the names of the "potential bloomers," providing evidence that the teachers attended to the bias presentation.

The major hypothesis, similar to the hypothesis tested in the Rosenthal and Jacobson (1968) experiment, that pupils who were the object of the expectancy bias would show greater pre- to posttest IQ improvement when compared to the remaining pupils, was tested with a three-factor analysis of covariance, using the pretest IQ as the covariate to control for initial IQ differences. Lord (1962) points out that the use of simple difference scores such as pre- to posttest IQ differences, may result in distortion due to uncontrolled regression effects. It also requires more rigid assumptions about the linear and ratio nature of the variable scale. Covariance used in randomly assigned groups may minimize some of these problems (see Evans & Anastasio, 1968). However, where groups are not randomly assigned, or where other assumptions are not met, it may not

[2] A more detailed description of the rating system, category definition, rater training, etc. can be found in the author's dissertation available from University Microfilms, Ann Arbor, Michigan 48106 (Document No. 69-8619).

TABLE 2
MEAN PRETEST FULL SCALE IQ FOR ALL GROUPS

Group	School 1			School 2			School 3			All schools	
	IQ	*n*	*SD*	IQ	*n*	*SD*	IQ	*n*	*SD*	IQ	*n*
Bias—Rated	108.9	19		117.2	19		109.9	23		111.9	61
Special	110.2	4	22.0	117.8	4	19.5	109.8	5	11.7	112.4	13
Nonspecial	108.6	15	10.9	117.0	15	13.6	109.9	18	15.1	111.7	48
Bias—Unrated	111.0	20		96.8	23		109.8	21		105.5	64
Special	114.0	4	15.5	93.4	5	9.3	113.5	4	15.8	105.9	13
Nonspecial	110.3	16	10.8	97.8	18	10.4	108.9	17	15.4	105.4	51
No Bias—Rated	118.4	20	15.6	99.5	21	12.7	114.8	23	15.1	110.9	64
No Bias—Unrated	101.1	19	12.4	92.4	16	5.2	113.8	22	20.0	103.6	57
Bias	110.0	39	12.2	106.0	42	15.8	109.8	44	14.5	108.3	125
No Bias	110.0	39	16.5	96.5	37	10.7	114.3	45	17.5	107.5	121
Rated	113.8	39	15.1	108.0	40	16.1	112.4	46	14.7	111.4	125
Unrated	106.1	39	12.8	95.0	39	8.7	111.8	43	17.8	104.5	121
Bias Special	112.1	8	17.7	104.2	9	18.7	111.4	9	12.9	109.1	26
Bias Nonspecial	109.4	31	10.7	106.5	33	15.3	109.4	35	15.0	108.4	99
All pupils	109.9	78		101.7	79		112.1	89		108.2	246

be possible to determine the appropriate adjustment for initial differences. Finally Lord points out that the use of a poor measure to adjust for initial differences is of negligible value. The reliability coefficients reported for the TOGA are sufficiently high (Flanagan, 1960) to reduce concern regarding the use of the IQ pretest score as a covariance adjustment.

IQ changes in the special and nonspecial pupils within the bias classes were compared, blocking on Rated and Unrated classrooms and on schools. The test for the hypothesis yielded an F of 2.12 (df = 1/101), which was not significant, indicating no effect as measured by the IQ change for the pupils who were designated as "bloomers" when compared to the remaining pupils in the class. Similarly, there were no significant differences when IQ subtest scores were compared. These findings are in contrast with those reported in Rosenthal and Jacobson (1968). The mean pre- and posttest IQ scores for the various groups are presented in Tables 2 and 3.

As can be seen from the tables, there

TABLE 3
MEAN POSTTEST FULL SCALE IQ FOR ALL GROUPS

Group	School 1			School 2			School 3			All schools	
	IQ	*n*	*SD*	IQ	*n*	*SD*	IQ	*n*	*SD*	IQ	*n*
Bias—Rated	114.4	19		133.9	19		123.7	23		124.0	61
Special	109.5	4	16.1	142.0	4	14.6	121.4	5	19.8	124.1	13
Nonspecial	115.7	15	15.1	131.7	15	13.2	124.3	18	18.3	123.9	48
Bias-Unrated	120.8	20		103.3	23		125.1	21		115.9	64
Special	117.0	4	22.4	97.6	5	8.6	128.5	4	5.7	113.1	13
Nonspecial	121.8	16	10.6	104.9	18	12.3	124.3	17	21.7	116.7	51
No Bias—Rated	131.9	20	14.4	107.6	21	13.7	132.7	23	20.1	124.2	64
No Bias—Unrated	106.3	19	18.2	102.7	16	14.0	124.4	22	23.5	112.3	57
Bias	117.7	39	14.3	117.1	42	20.0	124.3	44	18.7	119.8	125
No Bias	119.4	39	20.7	105.5	37	13.8	128.7	45	22.0	118.6	121
Rated	123.4	39	17.0	120.0	40	19.0	128.2	46	19.5	124.1	125
Unrated	113.7	39	17.3	103.1	39	12.6	124.7	43	21.4	114.2	121
Bias Special	113.2	8	18.5	117.3	9	25.8	124.6	9	14.9	118.6	26
Bias Nonspecial	118.9	31	13.1	117.1	33	18.4	124.3	35	19.7	120.2	99
All pupils	118.6	78		111.7	79		126.5	89		119.2	246

was considerable variability between classes in initial and follow-up IQ. Comparison of pretest IQs for critical groups (e.g., the between comparison of the special Bias pupils with pupils in No Bias classes and the within comparison of the special Bias pupils with the nonspecial Bias pupils) shows that randomization yielded special and nonspecial groups within schools of sufficiently similar pretest IQs.

Examination of Table 4 shows a substantial practice effect or gain from pre- to posttesting. The mean full scale gain for all pupils was 11 points, ($t = 14.86$, $df = 245$, $p < .0001$). The gains reported across schools and classes were not consistent and no simple explanation for these differences is apparent.

Further analyses nominally evaluated hypotheses relating to cross-class comparisons, such as the prediction that special children in the Bias classes would show more gains than the children in the No Bias classes; that children designated as nonspecial in the Bias classrooms would show less IQ gain than the children in the No Bias classrooms; and that children in the Rated classrooms would show greater pre- to posttest IQ gains than would the children in the Unrated classrooms. Only the last of these three hypotheses was supported ($F = 8.39$, $df = 1/197$, $p < .005$). Since the classes comprising this test were not randomly assigned, factors relating to assignment, observation or their interaction may be the source of the results.

The other major set of hypotheses dealt with predictions related to teacher-pupil interactions. The hypotheses were tested with multivariate analyses of variance (Cooley & Lohnes, 1962). The hypotheses that within the Bias Rated classes, teacher-pupil interaction ratings would change more with the special children than with the nonspecial children, and that across classrooms, teacher-pupil interactions with special pupils in the Bias classrooms would show a differential change when compared to teacher-pupil interactions with special pupils in the No Bias classrooms were not supported. There was, however, weak evidence of a differential change in the teacher-pupil interactions with the nonspecial pupils in the Bias classrooms when compared with the nonspecial pupils in the No Bias classrooms ($F = 3.60$, $df = 8/95$, $p < .01$). The contributions of the various teacher-pupil interaction variables were not predicted a priori, and are not easily or clearly related to teaching practice.

TABLE 4
ADJUSTED MEAN FULL SCALE IQ GAINS

Group	School 1	School 2	School 3	All schools
Bias Special pupils	1.63	13.74	10.70	8.69
Rated	−.60	25.83	5.62	10.26
Unrated	3.86	1.64	15.77	7.09
Bias Nonspecial pupils	9.39	10.50	16.97	12.29
Rated	7.02	16.10	16.13	13.08
Unrated	11.76	4.89	17.81	11.49
Bias Rated classes	3.21	20.97	10.88	11.69
Bias Unrated classes	7.81	3.27	16.79	9.29
No Bias Rated classes	12.00	8.19	13.31	11.17
No Bias Unrated classes	4.68	12.21	7.79	8.23
All Bias classes	5.51	12.12	13.83	10.49
All No Bias classes	8.34	10.20	10.55	9.66
All classes	6.93	11.16	12.19	10.09

DISCUSSION

The major hypotheses were not supported. There were, however, certain crucial differences in procedure which distinguish this finding from those presented by other authors. First, however, a noting of similarities is in order. The IQ test and the bias statements which accompanied the "test results" were exactly the same as those used by Rosenthal and Jacobson (1968). Approximately the same percentage of designated "bloomers" was chosen from each class. Two major differences between the present procedure and the Rosenthal and Jacobson study do exist: (*a*) the bias in the present study was introduced about 1 month into the second semester of the school year, presumably well after the teacher had formed impressions of her pupils; (*b*) retesting followed 2 months after the introduction of the

bias. In the Rosenthal and Jacobson study, the bias was introduced at the beginning of the school year, and retesting for IQ change was performed at the end of each semester. Their teachers, presumably, had not had time to form stable impressions before the introduction of the bias; similarly, the duration of the experiment was several months longer than in the present study. In the study by Conn, Edwards, Rosenthal, and Crowne (1968) the bias was introduced at the beginning of the second semester and pupils were retested at the end of that semester. Anderson and Rosenthal (cited in Rosenthal & Jacobson, 1968, p. 145 f) observed expectancy bias effects within the summer camp season. Likewise, Beez (1968) found bias effects within an 8-week period. Considering these studies together, it appears that neither the duration of the experiment nor the nature of the teacher's prior impressions have been shown to be critical variables.

Rosenthal (1969) combines 11 studies dealing with expectancy effects in educational settings. He computes a directional standard normal deviate for the "comparable" findings. By considering all the findings as essentially similar, he computes a joint one-tailed probability of .00033 supporting the position that expectancy effects have been reliably demonstrated. However, examination of the studies which he has combined shows that the findings were not at all similar and certainly not directly comparable. For example, while the Claiborn (1968) study, reported here, was based on approximately 2 months of "bias effects" for first graders, the comparison presented from Rosenthal and Jacobson (1968) apparently refers to the results of all classes after 1 year. The most similar comparison would seem to be the one-semester results of Rosenthal and Jacobson for first graders. Had that been used, substantially different standard normal deviates would be in the table. Similar arguments can be made for and against the inclusion of the nine other studies, and it does little to the strength of the expectancy-effect position to juxtapose substantially incongruent and dissimilar findings and represent them as parallel.

A major difference between the Rosenthal and Jacobson (1968) and the present study is the level of pretest IQ. Generally, in the present study, the pretest IQ was substantially higher than that reported in Rosenthal and Jacobson. In both cases, the TOGA norms proved to be inadequate resulting in added error variance.

There are two major conclusions to be drawn from the present study: (*a*) Further research needs to be conducted before the conclusions of the Rosenthal and Jacobson experiments become accepted as psychological fact. It should be clear from this short paper that at least one study which was sufficiently similar to the original paradigm has produced results which do not support, nor suggest that there is, an expectancy effect. (*b*) It would appear that the assessment of teacher-pupil interaction variables in terms of a relatively easily rated set of behavioral interactions has yet to prove its usefulness. Since the hypotheses relating to IQ change were not supported, little can be said about the ability of the rating procedures to capture teacher changes which were a result of the bias. There was some evidence that the presence or absence of the expectancy bias was related to the teaching behavior toward those children who were *not* included in the bias statement. However, the nature of the relationship and the configuration of the variable weights were not predicted and have little face validity. A more tenable general hypothesis is that as a result of biased expectations, some teachers changed their behavior but that these behavior changes cannot accurately or adequately be assessed by analysis in terms of identical changes for all variables for all *Ss*.

From this study, it appears that teacher behavior is moderately resistant to the kinds of bias or expectancy statements which make up much of our standardized testing programs. The evidence concerning the effects of giving teachers information about the abilities of their pupils on the pupils' academic performance remains equivocal. Considering the discussion presented in this paper and by others (Barber & Silver, 1968), caution should be

used in accepting significance levels and verbal conclusions on face value. Data can be combined in ways designed to maximize desired outcome and to capitalize on chance or other factors (as exemplified by the use of uncontrolled difference scores).

Unlike much research in education and psychology, the Rosenthal and Jacobson (1968) report has already begun to have an effect on educational practice. Of course, even if the teacher expectancy effects are accepted without qualification, the issue as to the magnitude of these effects (or amount of variance accounted for) has not been discussed. It is essential that psychologists be particularly careful in making conclusions from ambiguous data.

REFERENCES

BARBER, T., & SILVER, M. Fact, fiction and the experimenter bias effect. *Psychological Bulletin*, 1968, **70**, (6, Pt. 2).

BEEZ, W. Influence of biased psychological reports on teacher behavior and pupil performance. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, **3**, 605–606.

CLAIBORN, W. An investigation of the relationship between teacher expectancy, teacher behavior and pupil performance. (Doctoral dissertation, Syracuse University) Ann Arbor, Mich.: University Microfilms, 1968, No. 69-8619.

CONN, L., EDWARDS, C., ROSENTHAL, R., & CROWNE, D. Emotion perception and response to teacher expectancy in elementary school children. *Psychological Reports*, 1968, **22**, 27–34.

COOLEY, W., & LOHNES, P. *Multivariate procedures for the behavioral sciences*. New York: Wiley, 1962.

EVANS, S., & ANASTASIO, E. Misuse of analysis of covariance when treatment effect and covariate are confounded. *Psychological Bulletin*, 1968, **69**, 225–234.

FLANAGAN, J. *Tests of general ability, technical report*. Chicago: Science Research Associates, 1960.

LORD, F. Elementary models for measuring change. In C. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1962.

ROSENTHAL, R. The effect of the experimenter on the results of psychological research. In B. Maher (Ed.), *Progress in experimental personality research*, I. New York: Academic Press, 1964.

ROSENTHAL, R. *Experimenter effects in behavioral research*. New York: Appleton Century Crofts, 1966.

ROSENTHAL, R. Teacher expectation and pupil competence: Studies in the social psychology of self-fulfilling prophecies. Paper presented at the meeting of the American Association for the Advancement of Science, New York, December, 1967.

ROSENTHAL, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969, in press.

ROSENTHAL, R., & JACOBSON, L. *Pygmalion in the classroom*. New York: Holt, Rhinehat & Winston, 1968.

SHAMES, M., & ADAIR, J. Experimenter bias as a function of the type and structure of the task. Paper presented at the meeting of the Canadian Psychological Association, Ottawa, May 1967.

(Received December 1, 1968)