DOCUMENT RESUME

ED 059 247 TM 001 061

AUTHOR	Rosenthal, Robert; Rubin, Donald B.
TITLE	Pygmalion Reaffirmed.
INSTITUTION	Harvard Univ., Cambridge, Mass.
SPONS AGENCY	National Science Foundation, Washington, D.C. Div. of Social Sciences.
PUB DATE	Jul 71
NOTE	24p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Classroom Research; Educational Researchers; *Elementary Grades; *Expectation; Intelligence Quotient; Interaction Process Analysis; *Research Methodology; Statistical Analysis; *Student Teacher Relationship
IDENTIFIERS	*Hawthorne Effect

ABSTRACT

This study refutes the Elashoff and Snow (1970) critique of "Pygmalion in the Classroom," a study by Rosenthal and Jacobson (1968) on the effect of favorable teacher expectance on pupil achievement. Among the theses considered erroneous are: (1) That there is a wide variation in apparent results when different methods of data analysis are employed; (2) That the statistically significant effects of teacher expectation are dependent upon the choice of a particular method of data analysis; (3) That imbalance and doubtful randomization in the experimental and control groups invalidate the results of the analyses; and (4) That the study is isolated and unreplicated. (AG)

1. Overview: Pygmalion in the Classroom Reaffirmed

In this paper, an invited response to the critique of Rosenthal and Jacobson (1968) (RJ) given in Elashoff and Snow (1970) (ES), we demonstrate that the ES document in no way impugns the validity of the RJ experiment.

A central thesis in ES is that there was a "wide variation in apparent results" when different methods of data analysis were employed, and that the statistically significant effects of teacher expectation reported by RJ were dependent upon the choice of a particular method of data analysis. This thesis is seriously in error. Indeed, as we shall show, the net effect of the varied statistical analyses carried out in ES is greatly to increase the cross-method generality of the results reported by RJ.

A second thesis in ES is that "imbalance" and "doubtful randomization" in the experimental and control groups invalidate the results of the RJ analyses. As we shall demonstrate, there is absolutely no reason to doubt the validity of the results of RJ.

A third thesis in ES is that the RJ study is an isolated, unreplicated study. As will soon be clear, RJ is one of scores of studies indicating significant effects of interpersonal expectancy.

In addition, there are many other equally erroneous theses in ES to which we shall respond.

Before responding to ES in detail, we want to emphasize the basic simplicity of purpose and design of the RJ experiment. The intent was to study the effect of favorable teacher expectancy on pupil performance. The simplest experiment RJ might have done would have been to randomly assign some children to a condition of favorable teacher expectation and to retain the remaining children as controls. Because of the randomization, the average difference in posttest scores between the experimental and control group children would be an unbiased estimate of the effect of favorable teacher expectancy for the population for which the children are representative. In order to guide one's judgment as to whether the measured expectancy effect is real in the sense of replicable, some significance testing may often be desirable. To make such testing more powerful, that is, more able to detect real effects when they do, in fact, exist, we often try to control other sources of variation besides the treatment. Thus, the randomization in the RJ experiment was done within blocks of classrooms and a concomitant variable, the pretest, correlated with posttest, was recorded. Blocking and adjusting for individual differences on the pretest are procedures designed to increase the precision of the measurement of the expectancy effect or, equivalently, to increase the power of a test of the significance of the effect.

In what follows we shall demonstrate not only that the reanalyses in ES strongly support the conclusions of the RJ report but also that generally the criticisms offered in ES are unsound.

2. Additional Evidence for the Pygmalion Effect

It is consistently claimed in ES that wide differences in results arise when different dependent variables (posttest scores, gain scores, adjusted posttest scores) are employed and/or when the dependent variables are "transformed" (untransformed, renormed, truncated) and/or when various nonparametric methods are used. Despite the varied procedures employed in ES, the expectancy effects found in RJ remain undiminished.

Table 1 compares the RJ dependent variable (untransformed gain score) for total IQ with the eight other ES dependent variables for total IQ. (Unless further specified, references to IQ are to total IQ.) Within each grade level employed by ES, the RJ score and the 95% confidence interval for the RJ score are given along with the mean, median, lowest, and highest of the eight other scores. The means and medians of the ES scores agree remarkably well with the RJ scores. In addition, all ES scores fall well within the 95% confidence intervals for the RJ scores and thus are thoroughly consistent with them. In fact, the eight other ES scores are significant (two-tail, $p < .05$) if and only if the RJ score was significant (two-tail, $p < .05$) (note last column of Table 1). Clearly, then, these ES procedures reaffirm the validity of the RJ conclusions, and we are grateful to ES for the effort they expended in tabulating these additional dependent variables.

We are also grateful to ES for having redrawn one of the RJ figures, thereby suggesting our Figure 1. In their improvement over the RJ figure, ES tabulated the data non-cumulatively and showed the proportion of

Table 1

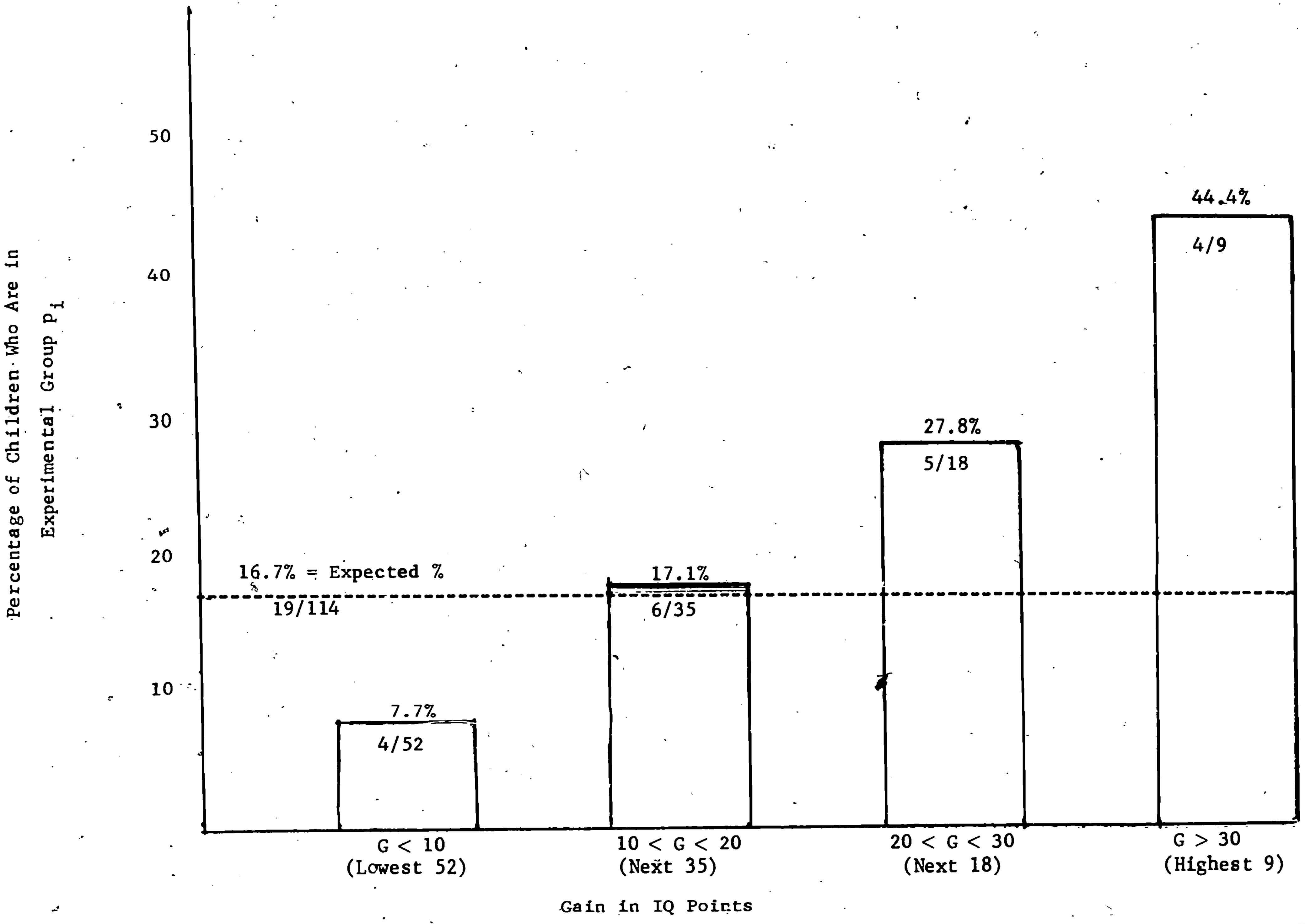
Comparison of Expectancy Advantage Scores in Total IQ Employed in RJ vs Eight Others Employed in ES

Grade[a]	RJ	95% Confidence Interval for RJ[b]	All Eight Other ES Scores				Total No. of Scores Sig. at $p < .05$, two-tail (Maximum possible = 9)
			Mean	Median	Lowest	Highest	
1 and 2	11.0	(4.7, 17.3)	11.6	10.7	9.2	15.9	9
3 and 4	1.8	(- 3.8, 7.4)	1.7	1.8	0.1	2.3	0
5 and 6	0.2	(- 6.2, 6.6)	1.1	0.1	-1.4	4.5	0

[a] Categories employed in ES.

[b] Based on Mean square within = 164.24.

Fig. 1. Proportion of Children Who Are Experimentals Gaining Various Amounts of Total IQ



children of grades one and two gaining varying amounts of IQ. However, they failed to note the statistical significance of the results they displayed. Based on ES' display of the data, Table 2 and Figure 1 show that there is a marked linear regression in the proportions (Snedecor and Cochran, 1967, pp. 246-247) of children who are experimentals on increasing levels of IQ gain ($\underline{p}$ = .0012, one-tail). (Unless otherwise specified, all subsequent $\underline{p}$ values are one-tail.) Thus, while less than 8% of the children gaining less than 10 IQ points are in the experimental group, over 44% of the children gaining 30 or more IQ points are in the experimental group. Assuming no effects of teacher expectation we would expect about 17% of the children in either of these categories to be in the experimental group. Table 3 and Figure 2 show the same analysis for posttest scores. Not surprisingly, the results indicate a similar linear trend, which is equally significant.

Another analysis comparing the proportions of experimental and control group children showing high posttest or gain scores is even more elementary. We employed the concept suggested in ES that, since there are 19 experimental children in the first two grades, the topmost 19 gain scores or posttest scores should be earned disproportionately often by the children of the experimental group. Seven of the top 19 gain scores were earned by children of the experimental group, more than twice as many as we would expect to find by chance ($\underline{p}$ < .02). Table 4 shows the results of this analysis and the results of the same analysis performed on posttest scores. As it turned out, the results were identical and hence significantly supported the expectancy hypothesis.

The similar analysis performed in ES was done within sex and classroom. Note that the top 19 children chosen by the ES method are not necessarily the top 19 children of the entire sample of 114 children from the first two grades. Their analyses yield $\underline{p}$ > .05 for gain scores and $\underline{p}$ < .001 for posttest scores. ES reported the nonsignificant result of their peculiar method of analysis, but failed altogether to mention the highly significant result it also yielded. Regrettably this failure to report th results of significance tests that do not support the null hypothesis is not an isolated instance, as we shall now indicate.

Table 2

Testing a Linear Regression of p_i on IQ Gain

	Gain in IQ Points				
Treatment	G < 10	10 < G < 20	20 < G < 30	G > 30	Total
Control (C)	48	29	13	5	95
Experimental (E)	4	6	5	4	19
Total (T)	52	35	18	9	114
p_i = E/T	.077	.171	.278	.444	.167

First order differences +.094 +.107 +.166

b = .112

S_b = .037

Z = 3.03

p = .0012, One-tail

Table 3

Testing a Linear Regression of p_i on IQ Posttest

Treatment	Lowest 52[a]	Next 35[a]	Next 18[a]	Highest 9[a]	Total
Control (C)	48	28	15	4	95
Experimental (E)	4	7	3	5	19
Total (T)	52	35	18	9	114
p_i = E/T	.077	.200	.167	.556	.167

First order differences +.123 -.033 +.389

b = .112

S_b = .037

Z = 3.03

p = .0012, One-tail

[a] Based on Ns given in ES Figure 2b.

Fig. 2 Proportion of Children Who Are Experimentals Showing Various Levels of Total IQ Posttest Scores

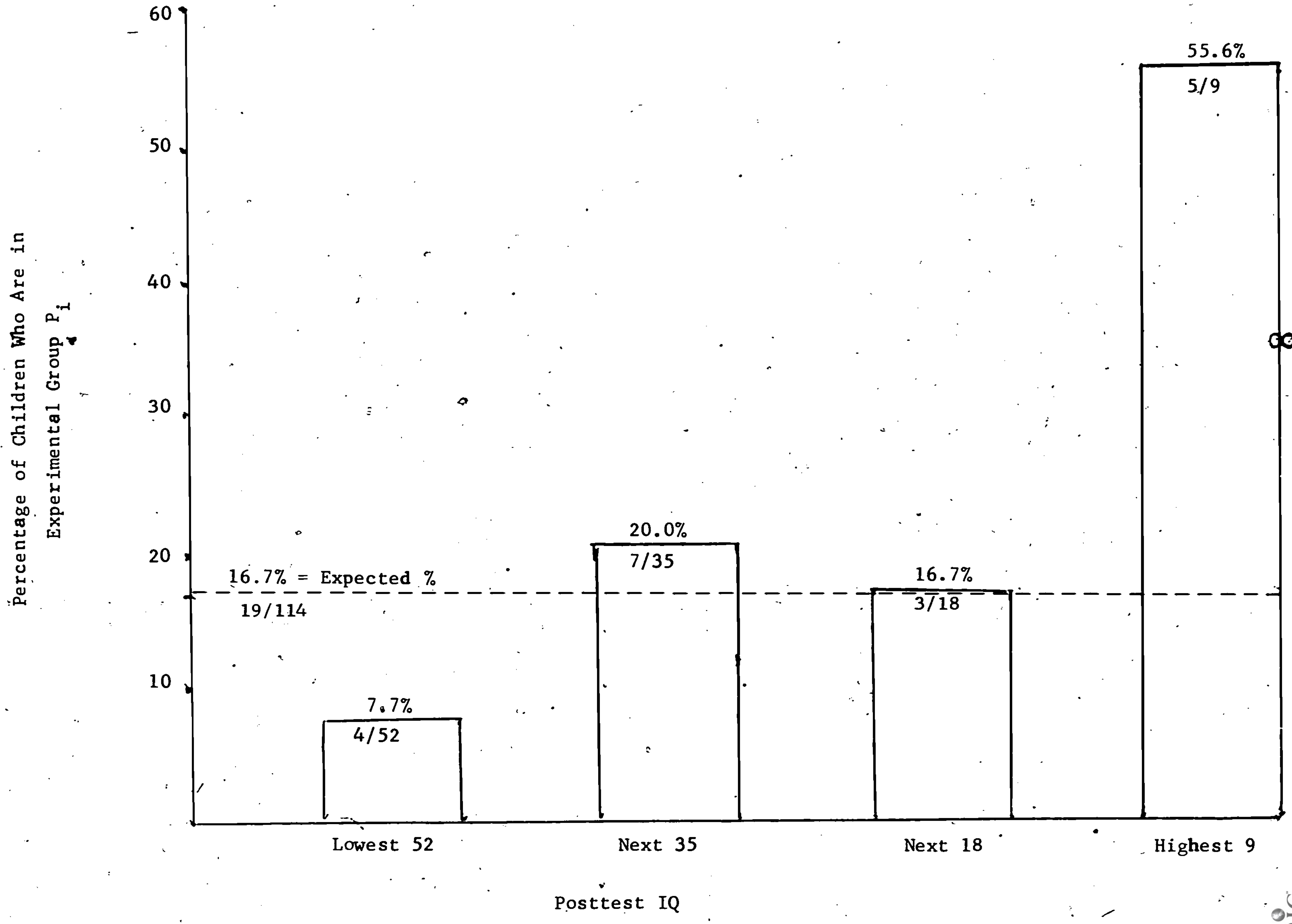


Table 4

Children Earning 19 Highest Scores
(Grades 1 and 2)

	Post Scores	Gain Scores
% of 19 Experimental Children (P_e)	37%	37%
% of 95 Control Children (P_c)	13%	13%
Difference	24%	24%
χ^2	5.05	5.05
Z	+ 2.25	+ 2.25
one-tail p	.0122	.0122
h[a]	0.57	0.57
Approximate Magnitude, (Cohen, 1969)	Medium	Medium

[a] h is defined as $(2 \arcsin\sqrt{P_e}) - (2 \arcsin\sqrt{P_c})$

Table 5

Percentage of Classrooms Showing Expectancy Advantage

		Posttest		Gain	
	Total N	%	One-tail p	%	One-tail p
Total IQ	17	76%	.025[b]	65%	.166[a]
Verbal IQ	18	61%	.240[b]	67%	.119[a]
Reasoning IQ	17	76%	.025[b]	88%	.001[b]

[a] Reported in ES

[b] Not reported in ES

In the discussion of the analysis by classrooms across all grades (ES tables 23 and 24), six low power (Cohen, 1969, pp. 35, 155) significance tests were performed on posttest scores and gain scores. Of these six, two were specifically mentioned in ES and both were nonsignificant. Of the remaining four not specifically mentioned, three were significant at $p < .025$; and all six were in the predicted direction (Table 5). Similarly, when examining raw gain scores for matched children, ES give a Wilcoxon signed-ranks test with $p < .05$, two-tail, which was subsequently discarded because of some mysterious "dubious validity," while a less powerful sign test found to be "nonsignificant" ($p = .059$) was not discarded. A similar kind of sweeping-under-the-rug of "undesirably" low p values was shown in the evaluation of Pygmalion by Jensen in his famous paper (1969, p. 107).

3. Initial Equivalence of Experimental and Control Groups

In summarizing the results of the previous section we emphasize that they strongly support the hypothesis of the positive effects of positive interpersonal expectation. Indeed ES seem to be aware of this fact since they repeatedly instruct their readers not to believe the results of their own analyses because of "doubtful randomization" and "imbalance" in the experimental conditions.

Imbalance in sample size has nothing to do with randomization or the ability to obtain unbiased estimates of the effects of teacher expectation. To claim that unequal sample sizes hopelessly confound the analysis of an experiment (ES pp. 28, 117) is to claim that a comparison of the means of two random samples is confounded if the sample sizes are not equal; this claim is clearly false.

In addition, there is no way in which the idea of "doubtful randomization" can be employed to impugn the validity of the Pygmalion experiment. In the first place, as RJ clearly pointed out, the children of the experimental condition were assigned to that condition at random; specifically, RJ used the table of random numbers provided by Arkin and Colton (1950). In the second place, when the analyses that have been performed on posttest and gain scores are performed on the pretest scores, they show no more difference between the experimental and control group children than would be expected by chance. Thus, for example, ES performed 36 overall _F_ tests of the significance of the difference between experimental and control group children on the pretest and obtained not a single _F_ significant at $p < .05$ (ES Tables 16, 17, 18). If we consider all interactions of treatment condition with other variables as well as main effects of treatment we find that 192 (nonindependent) _F_ tests of significance were made. Of these 192 _F_ tests, only one, a triple interaction, was significant at $p < .05$, a result that could easily have occurred by chance, yet was singled out for comment in ES. Similarly, when ES analyzed pretest differences between experimental and control group children employing classrooms as the sampling unit, they found no significant differences (ES Tables 23 and 24).

It may also be asked whether the linear regressions shown in Figures 1 and 2 and Tables 2 and 3 to be significant for IQ gain and IQ posttest might not also be significant for the pretest. Table 6 and Figure 3 show that this was not the case ($z < 1$).

Finally, employing the method of the "top 19" children introduced in ES, we can determine whether children of the experimental condition were overrepresented among the children earning the highest 19 scores on the total IQ pretest. Under the hypothesis of successful randomization we expect to find about three or four children of the experimental group among the top 19. What we find is just what we would expect under conditions of successful randomization: four of the top 19 were members of the experimental group ($X^2 = 0.05$, $df = 1$, $p = .82$).

In summary, since children were assigned to the experimental condition by means of a table of random numbers and since, furthermore, dozens of tests on the distribution of the pretest gave no indication that there had been any failure of randomization, it becomes most difficult to understand the continued concern shown in ES over "doubtful randomization." It must be concluded that ES' basis for not believing the effectiveness of randomization remains obscure and that the validity of the RJ experimental design has been thoroughly confirmed.

Table 6

Testing a Linear Regression of p_i on Pretest IQ

	Pretest Levels of IQ				
Treatment	Lowest 52[a]	Next 35[a]	Next 18[a]	Highest 9[a]	Total
Control (C)	44	31	13	7	95
Experimental (E)	8	4	5	2	19
Total (T)	52	35	18	9	114
p_i = E/T	.154	.114	.278	.222	.167
First order differences		-.040	+.164	-.156	

b = .035

S_b = .037

z = 0.95

p = .1711

[a] Based on Ns given in ES' Figure 2b.

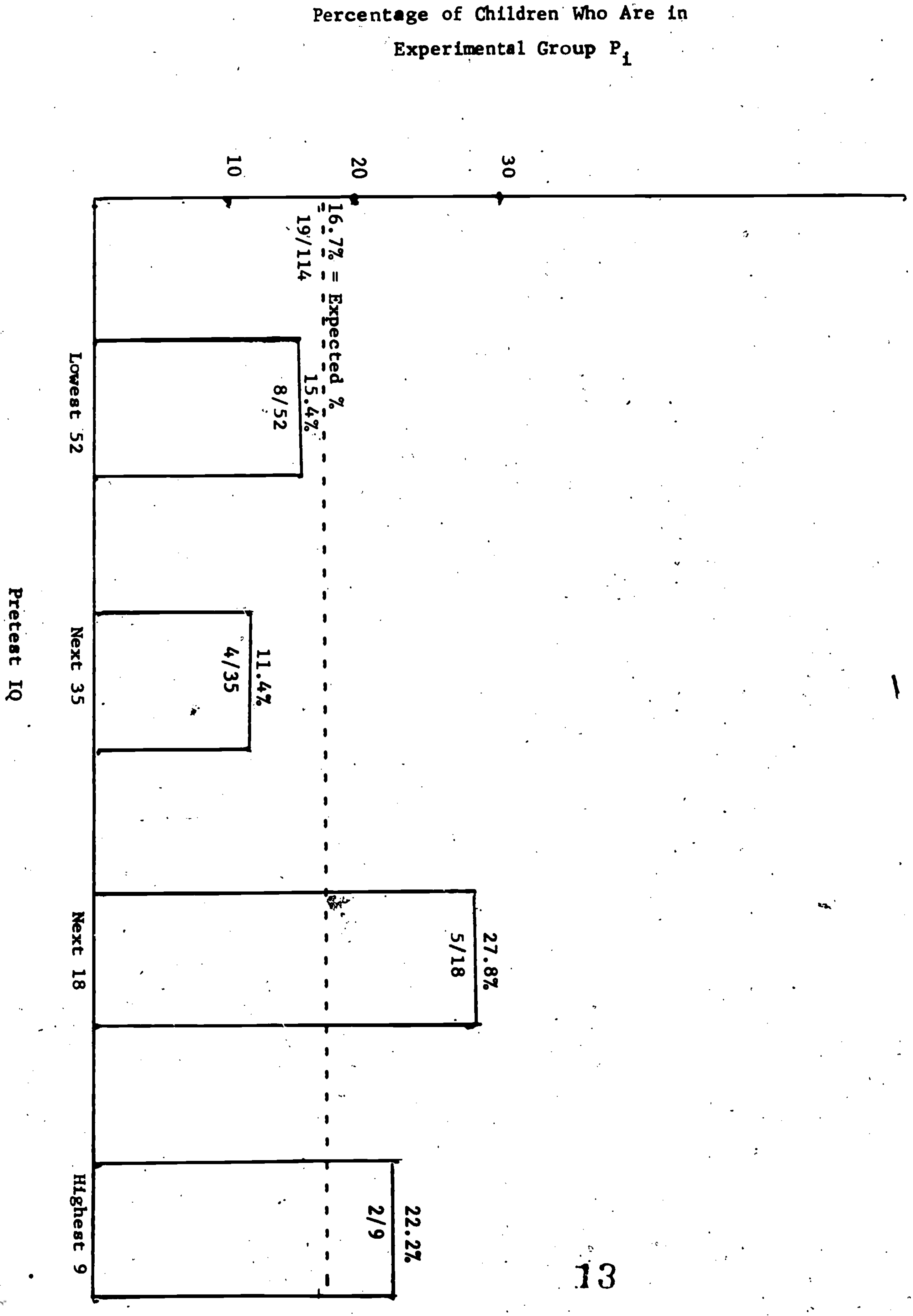


Fig. 3. Proportion of Children Who Are Experimentals Showing Various Levels of Total IQ Pretest Scores

4. Misleading Citation of Replication Research in ES

As mentioned earlier, the purpose of all the tests of significance performed is basically to evaluate the "reality" of the expectancy effect obtained, i.e., to determine its replicability under virtually identical conditions. There is another sense of replicability which has to do with the ability of the same or other investigators to obtain similar results. The latter kind of replication is of particular importance to the behavioral sciences as has been discussed elsewhere in detail (Rosenthal, 1966, 1969b).

In their brief mention of replications of the Pygmalion effect, only one study was mentioned by name, a failure to replicate by Claiborn (1969). In the doctoral dissertation upon which the Claiborn paper was based, it was candidly explained that two of the three teachers whose experimental condition was similar to that of the RJ study were either fully aware or partially aware of the nature and purpose of the experiment (Claiborn, 1968). Regrettably, in his subsequent article, Claiborn (1969) failed even to mention this difficulty in his discussion of his results. Interestingly, within his three classrooms similar to those in RJ, the tendency to obtain reversed results was strongly related to the teachers' degree of awareness of the purpose of the experiment.

Actually, at the time the Claiborn study appeared, numbers of studies showing significant positive effects of teacher expectation had been published and/or read at conventions (e.g., Beez, 1968; Burnham and Hartsough, 1968; Meichenbaum, Bowers, and Ross, 1969; Palardy, 1969). Therefore, the citation of only the Claiborn study is misleading.

Table 7 has been provided to give the reader an up-to-date picture of the results of studies of interpersonal expectation. Though many of the studies summarized are very recent, most of them have been summarized elsewhere (Rosenthal, 1969b, 1971). The first column shows for studies of teacher and counselor expectations the percentage yielding results at the .05, .01, and .001 levels of significance in either direction and the percentage yielding nonsignificant results. The second column gives

Table 7

Percentage of Studies Reaching Given p Levels

		Type of Study	
	Teachers	Experimenters	Total
Significance Level	(N=37)	(N=162)	(N=199)
$p \leq .05$ (one-tail)			
% in Predicted Direction	38%	33%	34%
% in Unpredicted Direction	0%	4^+%	4^-%
% Not Significant	62%	63%	63%
$p \leq .01$ (one-tail)			
% in Predicted Direction	14%	14%	14%
% in Unpredicted Direction	0%	1%	1%
% Not Significant	86%	85%	85%
$p \leq .001$ (one-tail)			
% in Predicted Direction	11%	9%	10%
% in Unpredicted Direction	0%	0%	0%
% Not Significant	89%	91%	90%

the corresponding data for studies conducted in lak·ratories rather than in everyday life situations. The percentages of studies reaching various levels of significance agree remarkably well, from studies of teachers to studies of experimenters. Considering those studies that are significant in the predicted direction vs those that are not, all three X^2s are less than one. It seems reasonable, then, to see both kinds of studies of interpersonal expectation as coming from a common population, and the third column of Table 7 shows the combined results. If there were no expectancy effect, we would expect to find about 10 studies of interpersonal expectation to have reached a $p \leq .05$ in the predicted direction; however, 67 studies have reached or exceeded that level, a virtually unobtainable result if there were no effect of interpersonal expectation.

In sum, the weight of the replicational evidence is very heavy, based as it is on the work of many investigators in many laboratories throughout the country. Although no experimental results in the behavioral sciences can be expected to show $p < .05$ in every study or even every other study, the ability of the effects of interpersonal expectancy to be demonstrated over a wide variety of dependent variables, investigators, laboratories, states, and even countries suggests a robustness not common to the ephemeral phenomena of the behavioral sciences. The Pygmalion effect is real.

5. Other Criticisms

Before going on to consider other criticisms in ES, we summarize very briefly what has been reported to this point:

(a) The *Pygmalion* effect does not depend upon the particular method of data analysis employed. This fact is clear using the evidence provided in ES.

(b) The experiment was fully randomized and there is no reason to doubt the initial equivalence of the experimental and control groups. This fact is clear using the evidence provided by approximately 200 tests performed in ES.

(c) *Pygmalion* is not an isolated study of interpersonal expectation. This fact is clear based upon results of scores of studies including many dealing specifically with teacher expectations.

In addition to the criticisms refuted above, there are other unsound criticisms of Pygmalion put forward in ES.

(1) ES imply that RJ should have employed stepwise regression in their analysis of a fully randomized experiment. At best, when all the appropriate interactions are entered, stepwise regression will give the same results as an analysis of variance. More usually, interactions and nonlinear trends are not entered, in which case stepwise regression eliminates important estimates and displays, and usually inflates the residual variance. In addition, stepwise regression inclines the user to assess the importance of a phenomenon using only the percentage of variance explained and to ignore not only the expected difference between means but even the direction of the effect (e.g. see ES Tables 11, 12, and 13).

(2) ES imply that RJ should have employed a rigid null hypothesis decision procedure. An ES imperative is to interpret no relation unless $\underline{p} < .05$ and to report $\underline{p}$ values less than .01 as < .01. The wisdom of this null hypothesis decision approach has been called into serious question not only by psychologists (Rosenthal, 1968; Rozeboom, 1960) but by a number of eminent statisticians as well. R. A. Fisher, for example, showed little patience with advice of the sort offered in ES, i.e., handy hints as to how and when to accept or reject hypotheses (Cochran, 1967). Fisher preferred to keep track of whatever $\underline{p}$ value was obtained and to wait and see what happened in subsequent observations. Finally, we find the ES orientation toward $\underline{p}$ values thoroughly inconsistent with the mental set they recommend, namely that of a detective rather than that of an attorney.

(3) ES imply that RJ's claim to increasing effects of teacher expectation in going from higher to lower grades is untenable. However, the RJ data showed a significant interaction of treatment with linear regression of grades ($\underline{t} = -2.69$, $\underline{df} = 308$, $\underline{p} < .01$, two-tail, Snedecor and Cochran, 1967, p. 278). In order to indicate the magnitude of this linear trend in average differences, we give the Pearson $\underline{r}$ between grade level and mean expectancy advantage per grade: $\underline{r} = -.86$ (RJ, p. 74). One display of this trend is shown in Table 8 and another in Figure 4. These results indicate that there is a clear and significant increasing effect of teacher expectation as one moves from higher to lower grades.

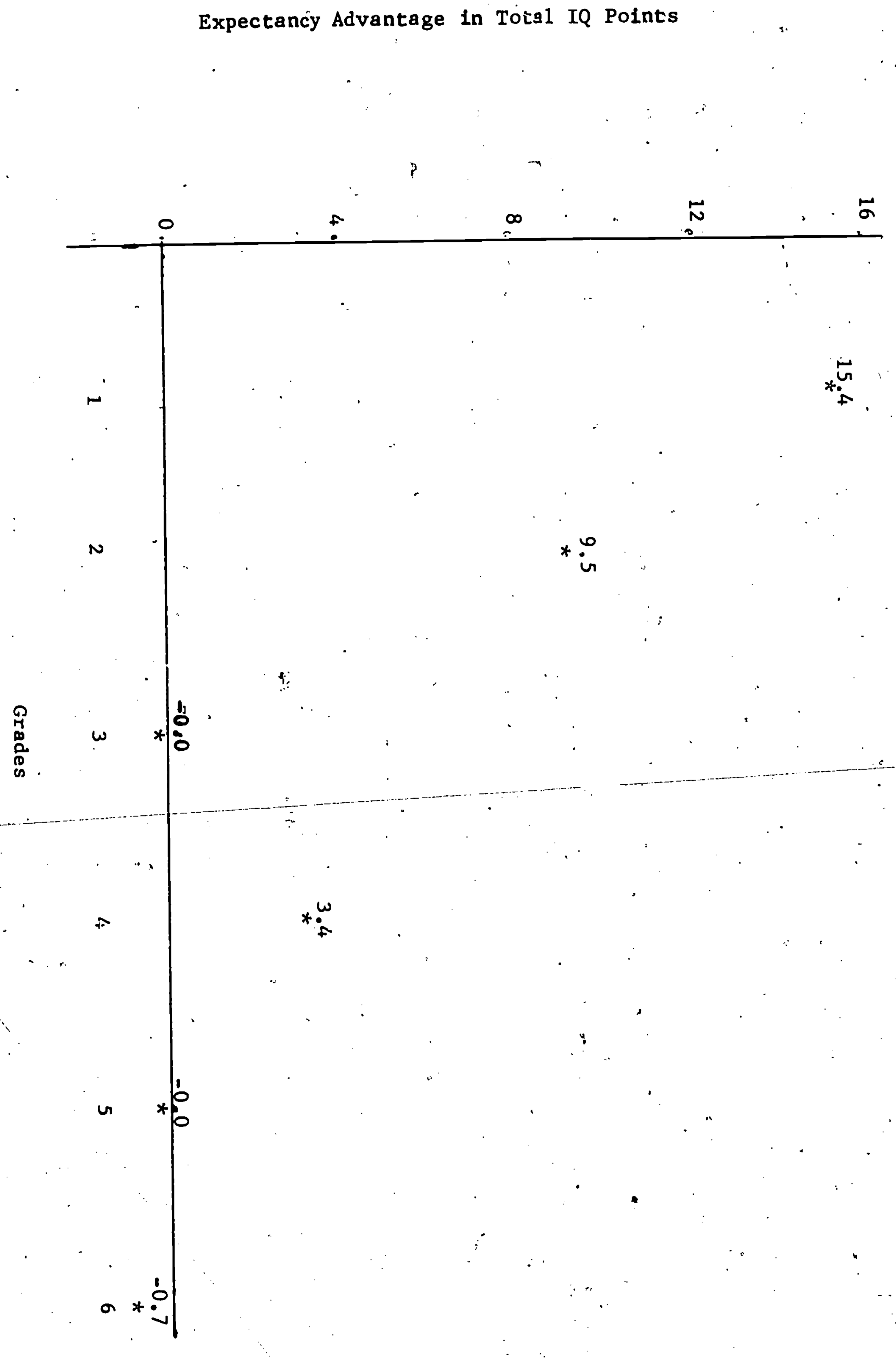


Fig. 4. Effects of Teacher Expectation in Six Grades

Table 8

Expectancy Advantage in Total IQ Gain After One Year

Grade	IQ Points	S.D. Units[a]	One-tail $p < .05$	Approximate Magnitude (based on Cohen, 1969)
1	+ 15.4	+ .83	.002	large
2	+ 9.5	+ .51	.02	medium
3	- 0.0	- .00	--	zero
4	+ 3.4	+ .18	--	small
5	- 0.0	- .00	--	zero
6	- 0.7	- .04	--	tiny (and reversed)
Total	+ 3.8	+ .21	.02	small

[a] σ = 18.48 based on pretest total IQ of all available children, N = 382.

Table 9

Expectancy Advantage in Reading Score Gain After One Year

Grade	Reading Scores	S.D. Units[a]	One-tail $p < .05$	Approximate Magnitude (based on Cohen, 1969)
1	+ .55	+ .56	.03	medium
2	+ .48	+ .48	.05	medium
3	+ .42	+ .42	.04	medium
4	+ .07	+ .07	--	very small
5	- .02	- .02	--	tiny (and reversed)
6	+ .08	+ .08	--	very small
Total	+ .17	+ .17	.05	small

[a] σ = 0.99 based on pretest reading scores of all available children, N = 313.

(4) ES imply that RJ should have employed the various ES data transformations. However, these transformations are statistically biased. Using the interval of 60-160, ES renormed by setting scores outside the range equal to the endpoints and truncated by discarding children outside the range. These would not have been biased procedures if they had been carried out only on pretest scores, even though they might restrict the generalizeability of the resultant analyses. On the other hand, when these procedures are applied to posttest scores they are biased and tend to diminish any real differences between the experimental and control groups. Specifically, if the experimental condition tends to increase or decrease scores, the above procedures would tend to distort or discard experimental scores more often than control scores thus making the means of the experimental and control groups more similar. For example, if one discarded all subjects whose posttests did not equal 100, the experimentals and controls would not differ on posttest.

(5) ES imply that RJ obtained effects that were trivial in magnitude even though they may have been significant statistically. In their discussion of the effects of teacher expectation on reading grades, ES point out that none of the differences in gains between experimental and control groups is as large as one full grade point equivalent (e.g., the difference between a grade of B and C). Since the standard deviation of the pretest reading grades was less than unity, ES appear to require an effect size to be larger than a standard deviation in order for it to be regarded as important. Such requirements, exceeding considerably a reasonable definition of even a large effect size (Cohen, 1969, p. 24) are very questionable. The actual effect sizes for reading scores are shown to be at least medium size in three of the six grades (Table 9). Note also the similarity of effect sizes between the expectancy advantages in reading score gains (Table 9) and total IQ gains (Table 8); the correlation between these two measures over the six grade levels is +.74 (RJ, p. 100).

(6) ES imply that the RJ dependent variables are unsuitable measures of intellectual performance. We find in ES a concern over the "low"

reliability ($\underline{r}$ = .74, Column 1, RJ Table A-30) of TOGA along with the implication that this threatens the validity of the Pygmalion experiment. Actually, unreliability (increased noise) can never account for the significant results of a fully randomized experiment; rather it can serve only to reduce power. We find also in ES the argument that the correlation of .65 between TOGA and subsequent ability track placement given in Rosenthal (1969a) does not adequately demonstrate validity. That correlation is higher than the correlation between scores on the nonverbal section of the Lorge-Thorndike and scores on the very same test retaken after an intervening summer. Finally, correlations between TOGA and other tests of intellectual performance are even higher (e.g., TOGA with Lorge-Thorndike: $\underline{r}$ = .73). Another recommendation found in ES is to employ raw scores as the dependent variable instead of IQ. We prefer to use IQ scores since it is IQs, not raw scores, that are used in the real world to make decisions.

(7) ES imply that RJ was insufficiently reviewed prior to publication. We feel, on the contrary, that RJ was unusually thoroughly reviewed prior to publication. In addition to having prior journal publication, the RJ research was solicited and approved for inclusion in a volume prepared for Division 9 of the American Psychological Association, Social Class, Race, and Psychological Development (Holt, Rinehart & Winston, 1968), edited by Martin Deutsch, Irwin Katz, and Arthur Jensen. Numerous other scholars in the behavioral sciences have requested permission to reprint the RJ research in their own volumes of readings, both before and after RJ was published. In addition, the award committee of Division 13 of the American Psychological Association presented the first prize of the Cattell Fund Award to the RJ research in 1967.

Given space, we could continue to refute criticisms of RJ and indicate many other errors in ES, but we feel that our point has been made.

6. Conclusions

We now conclude this response to the criticisms of Pygmalion (RJ) given in Elashoff and Snow (1970) (ES), having demonstrated the following:

(a) The results of the varied ES analyses are absolutely consistent with the results of the RJ analyses and indicate a significant effect of teacher expectations.

(b) RJ was a completely randomized experiment and the numerous ES tests of the success of the randomization give absolutely no reason to doubt the pre-experimental equivalence of the experimental and control groups and thus no reason to doubt the validity of the conclusions above.

(c) Positive effects of favorable interpersonal expectations have been obtained in numerous experiments conducted by dozens of researchers and thus the result in RJ can in no sense be considered a fluke.

(d) Although there were among the ES criticisms a few useful notions which we employed in this reply, in the main the numerous criticisms advanced in ES were neither sound nor constructive.

References

Arkin, H., and Colton, R. R. Tables for statisticians. New York: Barnes & Noble, 1950.

Beez, W. V. Influence of biased psychological reports on teacher behavior and pupil performance. Proceedings of the 76th Annual Convention of the American Psychological Association, 1968, 605-606.

Burnham, J. R., and Hartsough, D. M. Effect of experimenter's expectancies ("The Rosenthal effect") on children's ability to learn to swim. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May, 1968.

Claiborn, W. L. An investigation of the relationship between teacher expectancy, teacher behavior and pupil performance. Unpublished doctoral dissertation, Syracuse University, 1968.

Claiborn, W. L. Expectancy effects in the classroom: a failure to replicate, Journal of Educational Psychology, 1969, 60, 377-383.

Cochran, W. G. Footnote to an appreciation of R. A. Fisher. Science, 1967, 156, 1460-1462.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Elashoff, J. D., and Snow, R. E. A case study in statistical inference: Reconsideration of the Rosenthal-Jacobson data on teacher expectancy. Technical Report No. 15, Stanford Center for Research and Development in Teaching, School of Education, Stanford University, December, 1970.

Jensen, A. R. How much can we boost IQ and scholastic achievement? Harvard Educational Review, 1969, 39, 1-123.

Meichenbaum, D. H., Bowers, K.S., and Ross, R. R. A behavioral analysis of teacher expectancy effect. Journal of Personality and Social Psychology, 1969, 13, 306-316.

Palardy, J. M. What teachers believe -- What children achieve. Elementary School Journal, 1969, 69, 370-374.

Rosenthal, R. Experimenter expectancy and the reassuring nature of the null hypothesis decision procedure. Psychological Bulletin Monograph Supplement, 1968, 70, 30-47.

ERIC

Rosenthal, R. Empirical vs. decreed validation of clocks and tests. American Educational Research Journal, 1969, 6, 689-691. (a)

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal and R. L. Rosnow (Eds.) Artifact in Behavioral Research. New York: Academic Press, 1969. Pp. 181-277. (b)

Rosenthal, R. Teacher expectation and pupil learning. In R. D. Strom (Ed.) Teachers and the learning process. Englewood Cliffs, New Jersey: Prentice-Hall, 1971, Pp. 33-60.

Rosenthal, R., and Jacobson, L. Pygmalion in the classroom. New York: Holt, Rinehart and Winston, 1968.

Rosenthal, R., and R. L. Rosnow (Eds.) Artifact in behavioral research. New York: Academic Press, 1969.

Rozeboom, W. W. The fallacy of the null-hypothesis significance test. Psychological Bulletin, 1960, 57, 416-428.

Snedecor, G. W., and Cochran, W. G. Statistical methods. (6th ed.) Ames, Iowa: Iowa State University Press, 1967.