Journal of Educational Psychology
1984, Vol. 76, No. 1, 85–97

Magnitude of Teacher Expectancy Effects on Pupil IQ as a Function of the Credibility of Expectancy Induction: A Synthesis of Findings From 18 Experiments

Stephen W. Raudenbush
Graduate School of Education, Harvard University

This inquiry employs meta-analysis (Glass, 1977) to account for variability in the outcomes of experiments testing the effects of teacher expectancy on pupil IQ. The tenuous process of expectancy induction, wherein researchers supply teachers with information designed to elevate their expectancies for children actually selected at random, is viewed as the Achilles' heel of Pygmalion (Rosenthal & Jacobson, 1968) experiments. It was hypothesized that the better teachers know their pupils at the time of expectancy induction, the smaller the treatment effect would be. The data strongly supported this hypothesis. Hypotheses that the type of IQ test (group vs. individual) and type of test administrator (aware vs. blind to expectancy-inducing information) influence experimental results were not supported. The hypothesis that expectancy effects are larger for children in Grades 1 and 2 than for children in Grades 3–6 was supported. However, surprisingly, significant effects reappeared at Grade 7. Theoretical implications and questions for future meta-analytic research are discussed.

Fourteen years of research demonstrate the effects of teacher expectancy on a wide variety of outcomes but leave in doubt the question of expectancy's IQ effects (Baker & Crist, 1971; Brophy & Good, 1974; Rosenthal & Rubin, 1978; Smith, 1980). These IQ effects have been a subject of continuing and acrimonious controversy (Jensen, 1969, 1980; Miller, 1980; Rosenthal & Rubin, 1971; Thorndike, 1968). A problem underlying the dispute is that experimental tests of expectancy's effect on IQ have produced conflicting results (Kellaghan, Madaus & Airasian, 1982). When experimenters have reported significant effects, their critics have offered rival explanations that challenge the credibility of the findings (Pellegrini & Hicks, 1972; Thorndike, 1968). When experimenters have reported null effects, their critics, in turn, have speculated that methodological artifacts may have been the cause (Brophy & Good, 1974; Carter, 1970/1971; Rosenthal & Rubin, 1971). To date no study has sought systematically and formally to test those alternative explanations for variation in experimental findings. This inquiry employs meta-analysis (Glass, 1977; Light & Smith, 1971; Rosenthal & Rubin, 1982a) to conduct such a test.

Controversy Over Pygmalion

The first stage in the controversy focused on the original Pygmalion study (Rosenthal & Jacobson, 1968). In that now famous experiment, all of the predominantly poor children in the so-called Oak elementary school were administered a test pretentiously labeled the "Harvard Test of Inflected Acquisition." After explaining that this newly designed instrument had identified those children most likely to show dramatic intellectual growth during the coming year, the experimenters gave the names of these "bloomers" to the teachers. In truth, the test was a traditional IQ test and the "bloomers" were a randomly selected 20% of the student population. After retesting the children 8 months later, the experimenters reported that those predicted to bloom had

Anthony S. Bryk, Richard R. Light, and Sarah Lawrence Lightfoot deserve special thanks for their advice on this inquiry though responsibility for the conclusions is the author's.

A technical appendix is available to the interested reader at cost ($1.50) by writing the author.

Requests for reprints should be sent to Stephen W. Raudenbush, Harvard University, Graduate School of Education, 463 Gutman Library, Appian Way, Cambridge, MA 02138.

in fact gained significantly more in total IQ (nearly 4 points) and reasoning IQ (7 points) than the control group children. Further, at the end of the study, the teachers rated the experimental children as intellectually more curious, happier, better adjusted, and less in need of approval than their control group peers.

The ideological climate in 1968 was ripe with potential for both eager, noncritical acceptance and unusually harsh criticism of the study and its findings. Researchers, educators, and political activists were engaged in a heated debate over the causes of poor children's depressed school achievement. In the context of this explosive debate, relevant social scientific findings took on an added public significance.

In Ryan's (1971) widely read *Blaming the Victim*, Pygmalion served as a centerpiece in the attack on the belief that poor children's home backgrounds were the cause of their school problems. Kohl (1971, p. 84) argued that Pygmalion's findings "condemn the tracking system prevalent in elementary and secondary schools throughout the country." However, Thorndike (1968), Jensen (1969), and Elashoff and Snow (1971) claimed that because of its alleged methodological flaws the study provided no basis for the widely heralded findings. A classic methodological battle ensued (see Rosenthal and Rubin's reply, 1971).

Replication Efforts

Eventually debate over the original study subsided and the disputants focused instead on replication attempts. Baker and Crist (1971) reviewed 25 early studies and found that 11 of the 14 studies using teacher–pupil interaction as the outcome had found significant effects of expectancy. Of the 12 studies employing pupil achievement as the outcome, half reported significant effects. However, none of the nine studies using IQ as an outcome showed overall significant effects. The authors concluded that expectancy probably does not affect pupil IQ, a finding "supported by a background of decades of research suggesting the stability of human intelligence and its resistance to alterations by environmental manipulation" (p. 61).

More recently, Rosenthal and Rubin's (1978) review of 345 studies further substantiates the expectancy effect in a wide variety of settings including classrooms but leaves clouded the specific question of expectancy's IQ effects. Smith (1980) found that high-expectancy and control groups were separated, on average, by 0.38 of a common standard deviation across 78 estimates of expectancy's effect on achievement. However, the average effect size was only 0.16 for 22 estimates of expectancy's IQ effects.[1] She concluded that teacher expectancy had minimal effect on pupil ability.

In summary, interpersonal expectancy effects have been well documented across a wide variety of outcomes including teacher–pupil interaction and achievement. Yet the reported IQ effects, which stimulated the original controversy, remain in doubt, and speculations about the causes of success or failure in these studies remain untested.

Rationale and Hypotheses of the Present Inquiry

Timing of Expectancy Induction

Experiments assessing teacher expectancy effects actually involve two conceptually distinct phases. In the first phase researchers provide teachers with biased test scores or other information designed to elevate the teachers' expectancies for specific children who had actually been assigned at random to the experimental group. This is the "expectancy induction" phase.

The second phase of these experiments tests the expectancy hypothesis: Do teachers' expectancies, once altered by the expectancy induction, influence their behavior and subsequent pupil response? Clearly, if the induction fails there can be no treatment effect, not because of any failure in the theory, but simply because no treatment was implemented. Evidence from several early studies and commentary by

[1] Smith's (1980) meta-analysis consisted of 47 studies reporting 147 comparisons between experimental and control children. There were 22 comparisons or "effects" where IQ served as the outcome, based on 10 studies.

Rosenthal and Rubin (1971), Carter (1970/1971), and Brophy and Good (1974) suggest that the timing of expectancy induction may prove crucial to its success.

Five studies mentioned by Baker and Crist as using IQ as an outcome also measured the influence of the induced expectancies on teacher behavior. Close examination of three of these studies (Anderson & Rosenthal, 1968; Flowers, 1966; Kester, 1969) reveals the following pattern: (a) In each case expectancy induction occurred before teachers had had extensive contact with their pupils; (b) in each case teachers significantly differentiated their behavior toward experimental and control children; (c) in each case there was some evidence of IQ change.[2]

In contrast, in the two remaining studies (Claiborn, 1969; Jose & Cody, 1971): (a) Expectancy induction occurred during the spring of the school year after months of teacher–student contact; (b) experimenters reported no teacher behavior effects; and (c) there was no evidence of IQ change. Further, Jose and Cody (1971) provided evidence that teachers in their study had rejected the expectancy-inducing information. At the completion of that experiment, a survey revealed that 61% of the teachers "had not expected the children to show improvement as a result of the test information." Further,

> Others stated that they knew the children and their backgrounds and knew what the children could be expected to do.... The modification of expectancy *may have been too weak to overcome prior teacher expectancy based on other knowledge of the child.* (p. 47, italics added)

The notion that the timing of the expectancy induction influences its subsequent effects, although based on evidence from only five studies, is also consistent with cognitive dissonance theory (Festinger, 1957). Researchers have found that when persons are involuntarily exposed to new information, they tend to tune out, misperceive, invalidate, or forget information discrepant with ingrained beliefs or established patterns of behavior (Brock & Balloun, 1967; Ewing, 1942; Hastorf & Cantril, 1954; Kleinhesselink & Edwards, 1975; Wallen, 1942).

A major hypothesis of this inquiry, then, which emerges from an examination of early replication attempts and is consistent with dissonance theory, is that how well teachers know their pupils, as indicated by the amount of time spent with them in the classroom prior to expectancy induction, influences their susceptibility to persuasion by researcher-provided information about the children's intellectual potential. This hypothesis could not be tested directly because few experiments attempt to measure changes in teachers' beliefs or behavior. However, if the hypothesized expectancy effects on IQ are real, then the hypothesis concerning the timing of the expectancy induction leads to the prediction that the size of the expectancy effects will correlate negatively with the number of weeks of teacher–student contact predating expectancy induction.

Uncontrolled Measurement and Tester Bias

Critics of the original Pygmalion study emphasized allegedly poor measurement of IQ as a source of skepticism about its findings (Elashoff & Snow, 1971; Jensen, 1969; Thorndike, 1968). The statistically significant effects of teacher expectancy on pupil IQ depended on the rather dramatic gains reported for the experimental children in Grades 1 and 2. As Thorndike (1968) noted, very young children's scores on a paper-and-pencil test are especially sensitive to their willingness to try the test items. Perhaps as a result, the younger children's scores were both unexpectedly low and highly variable at the pretest. Thorndike suggested that the experimental treatment may have influenced children's motivation and that the group testing may have rendered the

[2] The subjects in the Anderson and Rosenthal (1968) study were retarded children. The "high expectancy" group suffered a *decline* in IQ, which may have resulted from the counselors' reported tendency to spend more time with apparently needier children. Flowers (1966) reported a treatment effect of 5.1 points in one school, but no apparent effect in a second school. Kester (1969), who employed gain score analysis, reported no effect of expectancy, but a more powerful analysis of covariance (in a reanalysis by the author) revealed a modest expectancy effect ($d = 0.27$, $p < .05$).

test scores unusually sensitive to such motivational effects.

Further, because in Pygmalion the teachers administered the posttests, Thorndike (1968) suggested that the reported IQ effects may have been caused by test coaching at the posttest. Pellegrini and Hicks (1972) designed their study specifically to test the hypothesis that expectancy effects operate through test coaching. Because expectancy effects were found when testers were aware of the researcher-provided designations of children's intellectual potential but were not found when the testers were blind to those designations, the authors concluded that the expectancy effects both in their study and in the original study had probably resulted from the test coaching of the experimental children.

The hypothesis that expectancy effects are made possible only by uncontrolled measurement or test coaching may be tested by data provided by replication attempts that varied in the conditions of their test administration. If expectancy effects can be found only as a result of poor measurement of IQ, more controlled measurement should make the IQ effects disappear. On the other hand, if the expectancy effect is real and fairly general, improved measurement may reveal even stronger, more statistically significant effects by reducing error variability.

Subjects' Age

The original Pygmalion suggested that younger children may be more susceptible to the influence of expectancy than older children. The present study sought to discover relationships between grade level of subjects and magnitude of treatment effect.

Method

Sample of Studies

Literature reviews (Baker & Crist, 1971; Brophy & Good, 1974; Jensen, 1980; Rosenthal, 1974) and an Educational Resources Information Center (ERIC) search produced 22 experimental studies of teacher expectancy where IQ or aptitude served as an outcome. From these sources, all studies employing IQ as an outcome and normal children in Grades 1–7 as subjects were included in the sample for the synthesis (n = 18).[3]

In all but one of the 18 studies, researchers randomly assigned children to treatments. One study (Flowers, 1966) compared two intact groups scoring similarly on the pretest: Experimenters assigned one group to be taught at a moderately higher ability level than the other for a year. Studies varied in ways directly relevant to the synthesis. The number of weeks of teacher–student contact predating the expectancy induction varied from 0 to 24. Tests were either group administered (n = 15) or individually administered (n = 3) by test administrators either aware (n = 10) or "blind" (n = 9) to the researcher-provided designations of the children's intellectual potential.[4] Studies also varied in the ages of their subjects. The sample is summarized in Table 1.

Other independent variables, less directly relevant to the major hypotheses of the inquiry, included statistical tests used, source of the study (doctoral dissertation or journal), year of publication, and a composite measure of methodological quality based on percentage of attrition, testing procedures, and statistical methods.

Analytic Procedure

Four methods devised to combine independent tests of a single hypothesis (Edgington, 1972; Fisher, 1938; Mosteller & Bush, 1954) enabled tests of the overall statistical significance of the expectancy effect for all 18 studies. The second task of the statistical analysis was to test hypotheses that the effects of the studies vary according to their treatments, methods, and subjects.

Statistical methods for assessing variations among findings responded to two different kinds of questions. First, do subgroups of studies differ significantly from each other in their findings? To answer this question, Rosenthal and Rubin's (1982a) statistical test of the significance of a contrast among study effect sizes was computed. Second, if it is found that two or more subgroups do differ significantly in their reported effects, do the combined results within any subgroup of studies indicate an experimental effect significantly greater than zero? To answer this type of question, two tests were employed: Mosteller and Bush's (1954) method for "adding zs," and Fisher's (1938) method for

---

[3] Of the studies mentioned in the literature reviews as employing IQ as an outcome, one (Goldsmith & Fry, 1970) could not be located. Listed by Baker and Crist (1971) as a forthcoming publication, that study reportedly involved high school students. Of the 21 studies located, three were eliminated from the analysis. One (H. Rosenthal, 1975) involved adult learners; a second (Anderson & Rosenthal, 1968) studied mentally retarded children; a third (Zanna, Sheras, & Cooper, 1975) referred to its outcome as a test of "aptitude" in three places but in two other places as a test of "achievement." Thus the final sample includes 18 studies, all of which employed IQ as an outcome and normal children in Grades 1–7 as subjects.

[4] Pellegrini and Hicks's (1972) study is counted twice because it included both an aware and a blind experimental condition.

Table 1
Description of Sample

Authors	Year of Publication	Estimated weeks of teacher–student contact prior to expectancy induction	Group vs. individual testing	Aware vs. blind test administrator	IQ test	Grades comparisons available	*d*	One-tailed *p*
Rosenthal, Baratz, & Hall	1974	2	Group	Aware	TOGA	1–6	0.02	.401
Conn, Edward, Rosenthal, & Crowne	1968	21	Group	Aware	TOGA	None	0.14	.206
Jose & Cody	1971	19	Group	Aware	TOGA	1–2 (Combined)	−0.03[a]	.791
Pellegrini & Hicks	1972	0	Group		PPVT	None	0.52	.010
Tester aware condition		0	Group	Aware		None	0.85	.003
Tester blind condition		0	Group	Blind		None	0.19	.242
Evans & Rosenthal	1969	3	Group	Aware	TOGA	None	−0.04	.709
Fielder, Cohen, & Feeney	1971	17	Group	Blind	TOGA	1–6	−0.02	.595
Fleming & Anttonen	1971	2	Group	Blind	KAIT	2	0.05	.224
Claiborn	1969	24	Group	Aware	TOGA	1	−0.13	.928
Kester	1969, 1972	0	Group	Aware	OLMA	7	0.27	.050
Maxwell	1970/1971	1	Individual	Blind	SB	2, 4	0.55	.002
Carter	1970/1971	0	Group	Blind	LT	7	0.30	.043
Flowers	1966	0	Group	Blind	OQS	7	0.18	.210
Keshock	1970/1971	1	Individual	Blind	SB	None	−0.01	.528
Henrikson	1970/1971	2	Individual	Blind	SIT	1	0.16	.250
Fine	1972	17	Group	Aware	CAT	2	−0.13	.877
Ginsburg	1970/1971	7	Group	Aware	CAT	1	−0.02	.519
Grieger	1970/1971	5	Group	Blind	CTMM-SF	1–4	−0.06	.637
Rosenthal & Jacobson	1968	1	Group	Aware	1–6	1–6	0.21	.016

Note. TOGA = Flanagan's Test of General Ability; PPVT = Peabody Picture Vocabulary Test; KAIT = Kuhlmann-Anderson Intelligence Test; OLMA = Otis Lennon Mental Abilities Test; SB = Stanford Binet; LT = Lorge Thorndike; OQS = Otis Quick Score; SIT = Slosson Intelligence Test; CAT = Cognitive Abilities Test; CTMM-SF = California Test of Mental Maturity—Short Form.

[a] This study reported an *F*-value without indicating the direction of the non-significant effect. The analyses were run twice under the assumption of a positive and a negative effect. The results were essentially identical.

"adding logs" (see Rosenthal, 1978, for details). Both tests were used because Rosenthal (1978) has reported that there is no uniformly best test and that the weaknesses of the two methods are to some degree offsetting.[5]

The main outcome variable for the meta-analysis was the effect size, denoted by the symbol d, which is the treatment effect in IQ points divided by the control group's standard deviation at the posttest. Note that in studies where the experimental children gain more than controls, d is positive, but when controls gain more than experimentals, d is negative. Effect sizes were based on total IQ scores and not on subtest scores.

Units of analysis depended on the research question. For some questions, each study (n = 18) served as the unit. For others, independent comparisons within grade levels served as units (n = 33).

Results

All Studies Combined

The effects sizes of the 18 studies, in standard deviation units, ranged from 0.55 to −0.13 (M = 0.11, SD = 0.20). Five of the studies achieved statistical significance, three at the .05 and two at the .01 level of significance. Of the remaining, nonsignificant effects, the experimental children scored higher than the control children in five cases, whereas the control children scored higher in eight cases. In no study, however, was the treatment effect both negative and statistically significant.

Three of four methods for conducting a combined significance test based on all 18 studies led to rejection of the null hypothesis of no effect of expectancy (see Table 2). Method 4, which weights studies by sample sizes, indicated no effect of expectancy. This result gave evidence that larger sample studies tended to produce smaller effects and that the analysis should take account of sample size as a potentially confounding variable.

Timing of Expectancy Induction

Evidence from early replication attempts had led to the hypothesis that the number of weeks of teacher–student contact prior to expectancy induction should correlate negatively with treatment effect sizes. A scatter plot (see Figure 1) showed the hypothesized association.

The correlation between effect size and weeks of prior contact was substantial (r = −.55). However, the plot shows an extreme curvilinear relationship. Because the correlation coefficient is a measure of *linear* association, it underestimates relationships that are actually curvilinear. One common way to facilitate study of curvilinear relationships is to transform the units in which the two variables are measured, that is to "linearize" the relationship. After employing a suitable transformation (Kirk, 1969),[6] a re-computed correlation coefficient (r = −.77), showed a strong negative association, suggesting that knowing the number of weeks of prior contact enables one to account for 59% of the variability in effect sizes.

A similar result is obtained by employing methods reported by Rosenthal and Rubin (1982a) for studying differences among study effect sizes. To test the hypothesis that effect sizes depend on weeks of prior contact, contrast weights were chosen to be inversely proportional to the number of weeks of prior contact.[7] The resulting statistical test supported the hypothesis (z = 2.75, p = .003). The total heterogeneity among the effect sizes may be measured by

Table 2
Results of Four Statistical Tests of the Effect of Expectancy on IQ Based on the Combined Results of 18 Experiments

Statistical test	Test statistic	p
Fisher (1938)	$\chi^2(36, N = 18) = 62.17$	.0025
Edgington (1972)	$\sum_1^N p_i = 6.84$	.04
Mosteller & Bush (1954)	$z = 2.12$	.017
Mosteller & Bush (1954)(weighted by degrees of freedom)	$z = 0.834$	*ns*

[5] The two tests produced highly convergent results. In Tables 3–6 the results are those of "adding zs."

[6] The reciprocal transformation (Kirk, 1969, p. 66) was employed here, after first adding a constant because both variables had values at or near zero: $x_{new} = -1/(2 + x_{old})$, $y_{new} = -1/(1 + y_{old})$. The negative reciprocal is used to preserve the variable's direction: Thus when x gets larger, $-1/(2 + x)$ also gets larger.

[7] Because some studies had a value of zero on prior contact and because denominators of zero should be avoided, a constant was first added to each study's value on that variable in creating these weights.

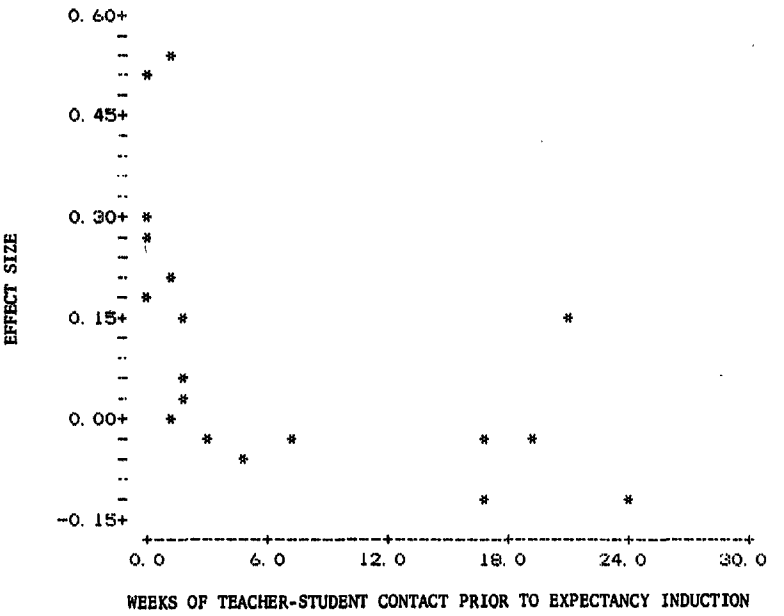

Figure 1. Scatter plot showing effect sizes of 18 experiments as a function of weeks of teacher–student contact prior to expectancy induction.

a χ^2 $(17, N = 18) = 14.65$. Because $z^2 = \chi^2$ $(1, N = 18) = 2.75^2 = 7.57$, this one degree of freedom contrast accounts for 52% (7.57/14.65 = .52) of the total heterogeneity of the effect sizes.

More specifically, four experiments having no prior contact yielded a mean effect size of 0.32. Three experiments having 1 week of prior contact obtained a mean effect size of 0.26. After more than 2 weeks of prior contact, the effect of expectancy seemed to disappear, as indicated by a mean effect size of −0.04 for those eight studies.

The combined significance test for all studies (see Table 2) had warned that sample size may have confounded the analysis. Indeed there was some indication of a relationship between weeks of prior contact and sample size ($r = .21$) and between sample size and effect size ($r = -.36$). Although the theoretical significance of sample size as a predictor of effect size is unclear, it did seem useful to adjust for its effect. New contrast weights, representing the component of prior contact orthogonal to sample size, were computed. The contrast based on these weights ($z = 2.56$, $p = .005$) suggested that the effect of prior contact was largely independent of sample size.[8]

To explore further the relationship between the timing of expectancy induction and experimental effects, studies were dichotomized into *high* (more than 2 weeks) and *low* (2 weeks or less) categories of prior contact. Each of four combined significance tests showed an expectancy significantly greater than zero for the low-contact studies. No method revealed a significant effect, in either direction, for the high-contact studies (see Table 3).

Uncontrolled Measurement and Tester Bias

To test the hypothesis that expectancy effects on IQ are made possible only by uncontrolled measurement, studies were broken down according to whether tests were individually or group administered (see Table 4). Though only three studies used individual tests, their results ($\bar{d} = 0.24$) were similar to results of studies using group tests ($\bar{d} = 0.22$) when considering only the low-contact studies.

[8] The contrasts for this test were the residuals obtained by regressing the original weights for prior contact (see Footnote 7) on sample size. These make appropriate weights because they sum to zero and represent the component of prior contact orthogonal to sample size.

Table 3
Results of Four Statistical Tests of the Effect of Expectancy on IQ Computed Separately for Studies Low and High in Teacher–Student Contact Prior to Expectancy Induction

	Prior contact			
	Low[a] (2 weeks or less)		High[b] (More than 2 weeks)	
Statistical test	Test statistic	p	Test statistic	p
Fisher (1938)	$\chi^2(20, N = 10) = 54.18$	<.001	$\chi^2(16, N = 8) = 7.98$	*ns*
Edgington (1972)	$\sum_{i=1}^{N} p_i = 1.57$	<.001	$\sum_{i=1}^{N} p_i = 5.26$	*ns*
Mosteller & Bush (1954)	$z = 4.16$	<.001	$z = -1.46$	*ns*
Mosteller & Bush (1954) (weighted by degrees of freedom)	$z = 2.52$	.006	$z = -0.827$	*ns*

[a] Mean effect size for 10 low-contact studies was 0.23.
[b] Mean effect size for eight high-contact studies was −0.06.

To test the hypothesis that expectancy's IQ effects are made possible only by test coaching (Pellegrini & Hicks, 1972; Thorndike, 1968), studies employing testers aware of the researcher-provided designations of children's intellectual potential were compared to studies employing testers blind to those designations.[9] The results indicated no main effect of tester awareness (z = −0.357), and no significant interaction between prior contact and awareness (z = 0.643). Among the "low contact" studies, both aware and blind conditions produced significantly positive effects of expectancy (see Table 5). These results suggest that the effect of prior teacher–student contact works independently of tester awareness of designations about children's intellectual potential.

Table 4
Mean Effect Sizes for Experiments Depending on Type of IQ Test (Group vs. Individual) and Extent of Teacher–Student Contact Prior to Expectancy Induction (High or Low)

	Prior contact			
	Low (2 weeks or or less)		High (More than 2 weeks)	
Test Administration	Size	n	Size	n
Group	0.22**	7	−0.06	8
Individual	0.24*	3		

* p <.05. ** p < .001.

This evidence is, however, correlational, and it conflicts with the interpretation of Pellegrini and Hicks (1972). Because in their study the contrast between aware and unaware conditions was created experimentally, and therefore lays a stronger basis for causal inference (Cooper, 1982), the null hypothesis of no effect of tester awareness should only be retained very tentatively.

Other direct or indirect measures of study quality—sophistication of statistical tests used, source of study (doctoral dissertation or journal), year of publication, and a composite measure of methodological quality (based on percentages of attrition, testing procedures, and statistical methods) were also employed as independent variables in the analysis. In each case, once prior teacher–student contact was accounted for, other information provided no help in understanding outcomes.

The Influence of Children's Age

As mentioned, the 18 experiments reported 33 independent comparisons between experimental and control children within grade levels. Mean effect sizes, broken down by three grade levels and two levels of prior contact, are shown in Table 6.

[9] Because Pellegrini and Hicks (1972) employed aware testers for a randomly selected half of the experimental children and blind testers for the other half, these two conditions are treated as separate studies in this section. Thus Table 5 shows 19 cases in all.

Table 5
Mean Effect Sizes for Experiments Depending on Test Administrator (Aware vs. Blind to Expectancy-Inducing Information) and Extent of Teacher–Student Contact Prior to Expectancy Induction

Test Administrator	Prior contact: Low (2 weeks or less) Size	*n*	Prior contact: High (More than 2 weeks) Size	*n*
Aware	0.34**	4	−0.035	6
Blind	0.21*	7	−0.135	2

* $p < .01$. ** $p < .001$.

The analysis with respect to grade level effects revealed the following findings:

1. Because controversy over Pygmalion had focused on effects across Gross 1–6, Grade 7 data were initially excluded from the analysis. The main effect of prior contact was again significant ($z = 1.87$, $p = .031$, one-tailed). The main effect of grade was nonsignificant ($z = 1.26$). The interaction of prior contact and grade was significant ($z = 1.85$, $p = .032$, one-tailed).

2. Considering all three grade levels, there was again a significant effect of prior contact ($z = 2.24$, $p = .013$, one-tailed). The apparent, though unanticipated, quadratic trend across grade levels within low contact (see Table 6) was suggestive, though nonsignificant ($z = 1.62$).

3. Expectancy effects were significantly greater than zero only for low-contact studies at Grades 1–2 and 7 (see Table 6).

Discussion

The central irony in teacher expectancy theory is that, despite its widespread dissemination as a paradigm for research and school reform, the proposition that made it famous has lived for 14 years under a cloud of controversy. Although a voluminous body of research (Rosenthal and Rubin, 1978; Smith, 1980) supports the expectancy hypothesis across a wide variety of outcomes, it was the original finding of the Pygmalion study that expectancy affected IQ, which inspired much of the ensuing debate. This inquiry sought to account for variability in the reported findings of experiments testing the effect of teacher expectancy on pupil IQ.

To summarize the main results, if we take all 18 experiments together, we find a small average effect of expectancy on IQ ($\bar{d} = 0.11$), which either achieves or fails to achieve statistical significance depending on the test employed and its underlying assumptions. This result is similar to that reported by Smith (1980; $\bar{d} = 0.16$).

The overall mean effect, however, certainly conceals more than it clarifies because these experiments appeared to differ markedly in their effectiveness in implementing the experimental treatment. Two of the first reported failures to replicate (Claiborn, 1969; Jose & Cody, 1971) provided reason to suspect that it may be difficult to persuade teachers to alter their expectations for children whom they already know on the basis of months of contact. The evidence supporting this assertion, based on all 18 experiments, is strong indeed. In fact more than half of the variability in the effect sizes of these experiments can be accounted for simply by knowing how many weeks of teacher–student classroom contact predated the expectancy induction. In interpreting this result, it is worth emphasizing that researchers who took pains to minimize the teacher's preexperimental contact with the children also appeared to take other measures to guarantee a strong treatment. Thus, for instance, Carter (1970/1971) manipulated not only IQ scores but also

Table 6
Mean Effect Sizes Depending on Grade Level of Subjects and Extent of Teacher–Student Contact Prior to Expectancy Induction

Grade	Prior contact: Low (2 weeks or less) Size	*n*	Prior contact: High (More than 2 weeks) Size	*n*
1, 2	0.31**	7	−0.09	8
3–6	0.04	9	0.03	6
7	0.25*	3		

* $p < .01$. ** $p < .001$.

achievement scores, past grades, and past teacher ratings, and also minimized the reactivity of the experiment by using the school system's customary testing program to obtain pre- and postmeasures. In Maxwell's (1970/1971) experiment, the principal met individually with teachers during the first week to emphasize the importance of "IQ results," which had, of course, been falsified to test for expectancy effects. It is possible then, that part of the strength of the relationship between prior contact and the size of the treatment effect is attributable to other features of strong treatment in the low-contact studies.

Practical Significance of the Expectancy Effect

An important result is the consistency of the treatment effect given little or no prior contact. The effect's robustness across methods of test administration strongly undermines the thesis that reported expectancy effects on IQ are made possible only by invalid measurement. But how large is the effect?

The practical significance of the effect is open to a variety of interpretations. Table 7 shows four measures of effect size for studies depending on the number of weeks of prior contact. The traditional measures r and r^2 are well known. Cohen's (1977) U_3 indicates the percentage of the experimental group that could be expected to outscore the control group given varying weeks of prior contact.

Table 7
Four Measures of Effect Size for 18 Experiments Depending on Weeks of Prior Teacher–Student Contact

Weeks of prior contact	Number of studies	$\bar{d}$[a]	r	r^2	U_3 (Cohen, 1977)
0	4	0.32	.158	.025	.626
1	3	0.26	.129	.017	.603
2	3	0.08	.040	.002	.532
>2	8	−0.04	−.020	.000	.492

Note. Assumes equal group sizes and equal within-groups variances. Under these conditions, $r^2 = d^2/(d^2 + 4)$.
[a] The rs, r^2s, and U_3 shown here are those associated with the $\bar{d}$ shown in this column.

Table 8
Mean Effect Sizes of 18 Experiments as Measured by Binomial Effect Size Display (Rosenthal & Rubin, 1982b) Depending on Weeks of Prior Teacher–Student Contact

Weeks of prior contact	% success experimentals	% success controls
0	58	42
1	56.5	43.5
2	52	48
>2	49	51

Note. This table assumes equal-sized groups. The expected percentages of experimental and control students who qualified are shown here if the median scores on an IQ test were the cutoff point for academic track placement.

A size of effect rather similar to that of Cohen's U_3 is conveyed by the binomial effect size display or BESD (Rosenthal and Rubin, 1982b). As indicated in Table 8, if the median score on the IQ test were used to select children for academic track placement, the experimental results for the four studies with no prior contact, for example, would lead us to expect that 58% of the experimental children, but only 42% of the control children, would qualify. This experimental advantage declines with each week of prior contact.

Expectancy Induction and Dissonance

The theory of cognitive dissonance (Festinger, 1957) provides a potentially powerful framework for explaining the variability, but not the existence, of the expectancy effect. That is, the theory provides the same plausible explanation for both the failure of expectancy induction in the high-contact studies and for its persuasiveness in the low-contact studies. In the former case, the theory leads us to expect that teachers will reject researcher-provided information discrepant with their own established patterns of interaction with children, thus nullifying any possibility of a treatment effect. In the latter case, it is the researcher-provided information, if sufficiently persuasive, that takes on the character of "prior knowledge" and may then shape the teacher's selective attention to the "new information": student

behaviors consonant and dissonant with that prior knowledge. It may be that the induced expectancies begin to have their effect by influencing the ways teachers interpret children's behavior during the first days or weeks of school.

Teachers have been found to be keen judges of pupil ability and are likely to reject test information that contradicts their personal knowledge (Fleming & Anttonen, 1971; Kellaghan et al., 1982). Because in Pygmalion experiments the expectancy-inducing information is false, this keen judgment, and not a need to reduce dissonance, may have led to the failure of expectancy induction in experiments high in prior contact. This explanation fails, however, to help us account for the success of expectancy induction in low-contact studies.

Expectancy Induction and Children's Age

The tendency, first reported in Pygmalion, for relatively strong expectancy effects to appear in the first and second grades, only to disappear in Grades 3–6 is supported by this synthesis. The speculation that this disappearance reflects children's decreasing vulnerability to the effects of adult influence, however, is flatly contradicted by the reappearance of the effect at Grade 7. This reappearance also contradicts the speculation that expectancy's reported IQ effects signify changes in young children's motivation in test taking and not their aptitude (Thorndike, 1968).

Why then do expectancy effects decrease during the elementary years? One plausible explanation is consistent with the finding that teachers' prior acquaintance with children immunizes them against expectancy induction. The Grade 7 studies (Carter, 1970/1971; Flowers, 1966; Kester & Letchworth, 1972) were among those that took care to prevent prior teacher knowledge of students from diluting the impact of the expectancy-inducing information. Further, their ability to do so may have been enhanced by the fact that the junior high school teachers in these studies had little or no contact with the children's previous teachers in elementary school.

Our "contact' variable accounts only for classroom contact. Yet it seems likely that once children have been in a school for a couple of years, their reputation if not their future teachers' personal knowledge of them may precede them each ensuing fall as they enter a new classroom. If so, the "expert" information provided by our researchers, designed to deceive teachers, may not have been deceptive at all for teachers in Grades 3–6 because it may have clashed with their own prior knowledge obtained outside the classroom.

Implications for Future Research

Do the results presented here hold up for non-IQ outcomes? Meta-analytic techniques used here may prove more powerful in accounting for variability in non-IQ outcomes because studies employing such outcomes are more numerous, adding degrees of freedom needed for a more powerful multivariate model to account for variability in study outcomes. The question of the timing of expectancy induction and variability in effects across grade levels should especially by investigated. Hypotheses concerning differential effects for children of varied backgrounds may be testable on a larger sample of studies.

The key methodological strength of research synthesized here is that its experimental character facilitates causal inference. Perhaps its key weakness is that assessments of artificially induced expectancy effects leave unanswered questions about the effects of naturally occurring expectancies. One strategy for preserving the experimental nature of the research while increasing ecological validity is to assess the impact of field interventions that enlist teachers' conscious involvement in altering expectancies. Evidence from several interventions (Greenfield, Banuazizi, & Gagnon, 1979; Kerman, 1979; Terry, 1977) encourages more investigation into the possibility that expectancies and their mediating mechanisms are subject to conscious change.

References

Anderson, D. F., & Rosenthal, R. (1968). Some effects of interpersonal expectancy and social interaction on institutionalized retarded children. *Proceedings of the 76th Annual Convention of the American Psychological Association, 3*, 479–480.

Baker, P. J., & Crist, J. L. (1971). Teacher expectancies: A review of the literature. In R. E. Snow & J. D. Elashoff (Eds.), *Pygmalion Reconsidered* (pp. 48–64). Worthington, OH: Charles A. Jones.

Brophy, J. E., & Good, T. L. (1974). *Teacher–student relationships: Causes and consequences.* New York: Holt, Rinehart & Winston.

Brock, T. C., & Balloun, J. L. (1967). Behavioral receptivity to dissonant information. *Journal of Personality and Social Psychology, 6,* 413–428.

Carter, D. L. (1971). The effect of teacher expectations on the self-esteem and academic performance of seventh grade students (Doctoral dissertation, University of Tennessee, 1970). *Dissertation Abstracts International, 31,* 4539-A. (University Microfilms No. 7107612)

Claiborn, W. (1969). Expectancy effects in the classroom: A failure to replicate. *Journal of Educational Psychology, 60,* 377–383.

Cohen, J. (1977). *Statistical power analysis for the social sciences* (rev. ed.). New York: Academic Press.

Conn, L. K., Edwards, C. N., Rosenthal, R., & Crowne, D. (1968). Perception of emotion and response to teachers' expectancy by elementary school children. *Psychological Reports, 22,* 27–34.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52,* 291–302.

Edgington, E. S. (1972). An additive model for combining probability values from independent experiments. *Journal of Psychology, 80,* 351–363.

Elashoff, J., & Snow, R. (1971). *Pygmalion reconsidered.* Worthington, OH: Charles A. Jones.

Evans, J., & Rosenthal, R. (1969). Interpersonal self-fulfilling prophecies: Further extrapolations from the laboratory to the classroom. *Proceedings of the 77th Annual Convention of the American Psychological Association, 4,* 371–372.

Ewing, T. A. (1942). A study of certain factors involved in changes of opinion. *Journal of Social Psychology, 16,* 63–88.

Festinger, L. (1957). *A theory of cognitive dissonance.* Evanston, IL: Row, Peterson.

Fielder, W. R., Cohen, R. D., & Feeney, S. (1971). An attempt to replicate the teacher expectancy effect. *Psychological Reports, 29,* 1223–1228.

Fine, L. (1972). The effects of positive teacher expectancy on the reading achievement of pupils in grade two (Doctoral dissertation, Temple University, 1972). *Dissertation Abstracts International, 33,* 1510-A. (University Microfilms No. 7227180)

Fisher, R. A. (1938). *Statistical methods for research workers.* London: Oliver & Boyd.

Fleming, E., & Anttonen, R. (1971). Teacher expectancy or my fair lady. *American Educational Research Journal, 8,* 241–252.

Flowers, C. E. (1966). Effects of an arbitrary accelerated group placement on the tested academic achievement of educationally disadvantaged students (Doctoral dissertation, Columbia University, 1966). *Dissertation Abstracts International, 27,* 991-A. (University Microfilms No. 6610288)

Ginsburg, R. E. (1971). An examination of the relationship between teacher expectations and student performance on a test of intellectual functioning (Doctoral dissertation, University of Utah, 1970). *Dissertation Abstracts International, 31,* 3337-A. (University Microfilms No. 710922)

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5,* 351–379.

Goldsmith, J. S., & Fry, S. (1970). *The test of a high expectancy prediction on reading achievement and IQ of students in grade ten.* Manuscript submitted for publication.

Greenfield, D., Banuazizi, A., & Gagnon, J. (1979). *An evaluation of the second year of Project STILE (Student–Teacher Interactive Learning Environment).* Unpublished report prepared for the Division of Education, Commonwealth of Massachusetts.

Grieger, R. M., II. (1971). The effects of teacher expectancies on the intelligence of students and the behaviors of teachers (Doctoral dissertation, Ohio State University, 1970). *Dissertation Abstracts International, 31,* 3338-A. (University Microfilms No. 710922)

Hastorf, A., & Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology, 49,* 129–134.

Henrikson, H. A. (1971). An investigation of the influence of teacher expectation upon the intellectual and academic performance of disadvantaged children (Doctoral dissertation, University of Illinois, Urbana, Champaign, 1970). *Dissertation Abstracts International, 31,* 6278-A. (University Microfilms No. 7114791)

Jensen, A. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1–123.

Jensen, A. (1980). *Bias in mental testing.* New York: Free Press.

Jose, J., & Cody, J. (1971). Teacher–pupil interaction as it relates to attempted changes in teacher expectancy of academic ability achievement. *American Educational Research Journal, 8,* 39–49.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1982). *The effects of standardized testing.* Boston: Kluwer-Nijoff.

Kerman, S. (1979, June). Teacher expectations and student achievement. *Phi Delta Kappan,* 716–718.

Keshock, J. D. (1971). An investigation of the effects of the expectancy phenomenon upon the intelligence, achievement, and motivation of inner-city elementary school children (Doctoral dissertation, Case Western Reserve, 1970). *Dissertation Abstracts International, 32,* 243-A. (University Microfilms No. 7119010)

Kester, S. W. (1969). The communication of teacher expectations and their effects on the achievement and attitudes of secondary school pupils (Doctoral dissertation, University of Oklahoma, 1969). *Dissertation Abstracts International, 30,* 1434-A. (University Microfilms No. 6917653)

Kester, S. W., & Letchworth, G. A. (1972). Communication of teacher expectations and their effects on achievement and attitudes of secondary school students. *Journal of Educational Research, 66,* 51–55.

Kirk, R. E. (1969). *Experimental design: Procedures*

for the behavioral sciences (2nd ed.). Belmont, CA: Brooks/Cole.

Kleinhesselink, R. R., & Edwards, R. E. (1975). Seeking and avoiding belief discrepant information as a function of its perceived refutability. *Journal of Personality and Social Psychology, 59*, 250–253.

Kohl, H. (1971). Great expectations. In R. E. Snow & J. D. Elashoff (Eds.), *Pygmalion reconsidered* (pp. 81–85). Worthington, OH: Charles A. Jones.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review, 41*, 429–471.

Maxwell, M. L. (1971). A study of the effects of teachers expectations on the IQ and academic performance of children (Doctoral dissertation, Case Western Reserve, 1970). *Dissertation Abstracts International, 31*, 3345-A. (University Microfilms No. 7101725)

Miller, H. L. (1980). Hard realities and soft social science. *Public Interest, 59*, 67–83.

Mosteller, F., & Bush, R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook on social psychology, Vol. I: Theory and Method.* (pp. 289–334). Cambridge, MA: Addison-Wesley.

Pellegrini, R., & Hicks, R. (1972). Prophecy effects and tutorial instruction for the disadvantaged child. *American Educational Research Journal, 9*, 413–419.

Rosenthal, R. (1974). *On the social psychology of the self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms* (Module 53). New York: MSS Modular Publications.

Rosenthal, H. (1975). The effect of teacher expectancy upon the achievement and intelligence scores of adult students. *Dissertation Abstracts International, 35*, 7017-7018.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185–193.

Rosenthal, R., Baratz, S., & Hall, C. M. (1974). Teacher behavior, teacher expectations, and gains in pupils' rated creativity. *Journal of Genetic Psychology, 124*, 115–121.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York: Holt, Rinehart & Winston.

Rosenthal, R., & Rubin, D. B. (1971). Pygmalion reaffirmed. In J. D. Elashoff & R. E. Snow (Eds.), *Pygmalion reconsidered.* (pp. 289–334). Worthington, OH: Charles A. Jones.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377–386.

Rosenthal, R., & Rubin, D. B. (1982a). Comparing effect sizes of independent studies. *Psychological Bulletin, 92*, 500–504.

Rosenthal, R., & Rubin, D. B. (1982b). A simple, general purpose display of experimental effect. *Journal of Educational Psychology, 74*, 166–169.

Ryan, W. (1971). *Blaming the victim.* New York: Pantheon.

Smith, M. L. (1980). Teacher expectations. *Evaluation in Education, 4*, 53.

Terry, J. T. (1977). *Student performance and school related attitudes as a function of teacher expectation and behavior.* Unpublished doctoral dissertation, Boston College.

Thorndike, R. L. (1968). Review of Pygmalion in the classroom. *American Educational Research Journal, 5*, 708–711.

Wallen, R. (1942). Ego-involvement as a determinant of selective forgetting. *Journal of Abnormal and Social Psychology, 37*, 20–39.

Zanna, M. P., Sheras, P. L., & Cooper, J. (1975). Pygmalion and Galatea: The interactive effects of teacher and student expectancies. *Journal of Experimental Social Psychology, 11*, 279–287.

Received December 8, 1982
Revision received March 25, 1983 ■