# Beleaguered *Pygmalion*: A History of the Controversy Over Claims That Teacher Expectancy Raises Intelligence

Herman H. Spitz
*Princeton, NJ, USA*

The 1968 publication of Rosenthal and Jacobson's *Pygmalion in the Classroom* offered the optimistic message that raising teachers' expectations of their pupils' potential would raise their pupils' intelligence. This claim was, and still is, endorsed by many psychologists and educators. The original study, along with the scores of attempted replications and the acrimonious controversy that followed it, is reviewed, and its consequences discussed.

> If animals become "brighter" when expected to by their experimenters, then it seemed reasonable to think that children might become brighter when expected to by their teachers. (Rosenthal & Jacobson, 1968a, p. 65)

Consider the implications of this remark, an almost identical version of which was repeated frequently thereafter, not only in subsequent publications but 95 pages later, and again 14 pages after that, with the quotation marks around "brighter" omitted. If, as seems likely, the quotation marks implied that animals (usually laboratory rats) really did not become "brighter," did the subsequent absence of quotation marks imply that the rats actually did become brighter? Or, as Fiske (1978) asked, "do these effects change organisms or only data?" (p. 394). Leaving aside the small percentage of fraudulent experimenters, as well as recording errors, Rosenthal ascribed the expectancy effect primarily to experimenters (frequently students) unwittingly treating the "brighter" rats differently than they treated the control group: handling them more, and more gently, giving cues in various ways, and so on. Likewise, experimenters who were testing animals that were expected to perform poorly "treated them in some subtle fashion such as to

Direct all correspondence to: Dr. Herman H. Spitz, 389 Terhune Road, Princeton, NJ 08540, USA. E-mail: hermspitz@aol.com

**INTELLIGENCE 27(3)**: 199–234
ISSN: 0160-2896

produce dull behavior" (Rosenthal, 1966, p. 177). In sum, according to Rosenthal the major effects of experimenters' expectancies on animal performance generally resulted from the experimenters' unwitting cues (the Clever Hans effect), which probably varied with the type of animal. But the rats could hardly have gained suddenly in rat intelligence, whatever that is.

Then what about the children? Is intelligence so malleable that teachers' expectancies can raise it? Did *Pygmalion in the Classroom* (henceforth *Pygmalion*) by Rosenthal and Jacobson (1968a) prove the wisdom of Rosenthal's analogy and did subsequent studies support it? This comprehensive review of studies and discussions concerned with whether teacher expectations can raise intelligence, measured by an intelligence test, is aimed at answering these questions.[1] With the 1992 reissue of Rosenthal and Jacobson's book and the latest exchange between Rosenthal (1994, 1995) and Snow (1995), the *Pygmalion* controversy has extended for three decades with no sign of a reconciliation, perhaps because it reflects the seemingly indestructible nature/nurture controversy and disputes about the malleability of intelligence.

## The Rosenthal–Jacobson Study and Its Immediate Sequelae: 1966–1971

In the mid-1960s, during a decade of promise (generated largely by Skinnerian behaviorism) that low intelligence need not be tolerated, Robert Rosenthal and Lenore Jacobson produced and disseminated widely the results of a study that was perhaps the most startling evidence of environmental power that had yet been presented. The early results of that study were described briefly in the book by Rosenthal (1966) on experimenter effects in behavioral research and more fully in a relatively unpretentious 4-page journal article that same year (Rosenthal & Jacobson, 1966). These were followed two years later by their article in *Scientific American* (Rosenthal & Jacobson, 1968b), their contributed chapter (Rosenthal & Jacobson, 1968c) and, to cap it all, their comprehensive book (Rosenthal & Jacobson, 1968a). Based on Rosenthal and Jacobson's findings, if teachers are told that tests indicate that certain pupils (secretly chosen at random) are very likely to show an academic spurt in the near future, mirabile dictu, many of those children will grow intellectually.

Experiments on expectancy effects should be set in perspective. The theoretical nucleus is a concept long studied by psychologists, sociologists and philosophers under various labels (Rosenthal, 1966; Gozali & Meyen, 1970; Zuroff & Rotter, 1985; Wineburg, 1987a) crystallized by Merton (1948) as the "self-fulfilling prophecy." Sometimes a concept wanders about without a single consensual anchor until it is given a compellingly appropriate name, and that is what happened with the self-fulfilling prophecy, although the term "expectancy effects," or some variant of it, continues to be used in psychology. The concept is simple enough: If we prophesy (expect) that something will happen, we behave (usually unconsciously) in a manner that will make it happen. We will, in other words, do what we can to realize our prophecy.

Rosenthal (1966) and others had for some time been presenting evidence that experimenters unwittingly influence the results of their research, but in *Pygmalion*, Rosenthal and Jacobson substituted teachers (and later physicians, therapists and employers) in place of experimenters. That is, instead of assessing the effects of *experimenters*' expectancies on the performance of human or animal subjects, Rosenthal and Jacobson assessed the effects of *teachers*' expectancies on the intelligence test scores of their pupils (Rosenthal, 1966). Their report of success created a sensation in the media (described in Elashoff & Snow, 1971; Wineburg, 1987a). For instance, *The New York Times* featured the results in a front page headline declaring: "Study Indicates Pupils Do Well When Teacher Is Told They Will" (Leo, 1967). Kohl (1968) in *The New York Review of Books* and Coles (1969) in *The New Yorker* wrote favorable reviews (both are reprinted in Elashoff & Snow, 1971). Approving commentary soon appeared in textbooks. Nevertheless, the reviews in professional journals were not uniformly rhapsodic; some were extremely critical, as we shall see, and a turbulent controversy developed.

Not only did *Pygmalion* kindle debate among psychologists, educators, and sociologists, it had broader social and educational consequences as well. In a Washington, D.C. case, *Hobson* v. *Hansen* (269 F. Supp. 401, 1967), *Pygmalion* was used by the plaintiffs in arguing their winning case to restrict the use of group tests for placing students in ability tracks. As a consequence the school system's ability-track program was abolished. The judge called the special track "an inflexible straitjacket . . . and he made much of the effect of labeling and stigmatization on both the children and the expectations of the teachers" (Elliott, 1987, p. 10). In the well-known case of *Larry P.* v. *Riles* (495 F. Supp. 926, 1979), which led to the State of California's prohibition on the use of intelligence tests for determining placement in EMR (educable mentally retarded) classes, plaintiffs also drew on *Pygmalion* in arguing their case, but the defense countered with witnesses who were very critical of the study. Apparently they made an impression on the judge who did not, "as judge Wright had done in *Hobson*, have much to say about it" (Elliott, 1987, p. 106). In *Bradley* v. *Milliken* (408 US 717, 1974), a Detroit case which led to the establishment of various procedures aimed at achieving better racial balance in the schools (and resulted in a massive exodus of whites to the suburbs), "Witnesses for both the defendants and plaintiffs were intoxicated with the popular expectancy hypothesis . . . The plaintiffs' expert witness on education laid great stress on the findings of a much-publicized book [*Pygmalion*] on the subject" (Wolf, 1981, p. 112), although they stressed the effects of teacher expectancy on academic performance rather than intelligence. *Pygmalion*, then, was not simply a scholarly exercise; it contributed to public policy deliberations and educational decisions. One would hope that the data were strong enough to support so portentous a role.

In 1992, *Pygmalion in the Classroom* was reissued by its original publisher as a College Edition, with the text unchanged and no acknowledgment that the study has some strong critics. Reportedly it has been selling briskly.

The experiment took place in the elementary school where the junior author, Lenore Jacobson, was principal.[2] The Oak School (pseudonym) was situated in a "somewhat run-down section of a middle-sized city" (Rosenthal & Jacobson, 1968b, p. 19), later revealed to be South San Francisco. About 17% of the students were Mexican, the only minority group. For each of grades 1 through 6 there were three classrooms because the school used an ability-tracking system that placed children in a slow, medium or fast classroom depending on whether their scholastic performance (mainly reading) was below average, average or above average.

In May 1964, the teachers at Oak School were asked to administer a test to all children in grades K through 5 (the pretest). Each teacher administered the test to his or her class. However, the teachers were not told the true name of the test or that it was an intelligence test, the Tests of General Ability (TOGA, Flanagan, 1960). Instead, they were told it was a test from Harvard University that predicted academic "blooming" or "spurting" by most of the pupils who performed well on the test. For the elementary grades there are three forms of the TOGA: one for grades K-2 (administered in this experiment to all K and grade 1 classes), another for grades 2–4 (administered to all grades 2 and 3 classes), and a third for grades 4–6 (administered to all grades 4 and 5 classes). Although the experimenters were interested (surreptitiously) in raising *intellectual* level, the teachers were apparently told to expect *academic* blooming.[3] To implement the charade, the cover of the TOGA was replaced by a cover with the impressive (albeit fanciful) name of the test from Harvard: Test of Inflected Acquisition. Additionally, each teacher was given an information sheet explaining that the primary interest of the Harvard study of inflected acquisition, which was supported by the National Science Foundation (*that* at least was true), was in children expected to "show an unusual forward spurt of academic progress ... within the next year or less" (Rosenthal & Jacobson, p. 66). At the same time, the study was said to provide a final validity check on the new test's ability to pick out these children. At pretest there were 305 children in the control group and 77 in the experimental group, and at that time the timetable for two of the three future testing sessions was divulged to the teachers. A final testing, scheduled for 2 years after the May 1964 pretesting, was not mentioned.

## A Brief Digression on the TOGA and IQ Tests in General

Developed as a group test of nonverbal intelligence for grades K to 12, the TOGA does not require the testee to read or write. There are two parts: Part 1 has items requiring information, vocabulary, and conceptual ability and yields a Verbal Score that Rosenthal and Jacobson referred to as Verbal IQ (35 items for K to 4, 45 items for higher grades); Part 2 has items requiring the ability to understand figural relationship and was designed to yield a noncultural abstract Reasoning Score that Rosenthal and Jacobson referred to as Reasoning IQ (28 items for K to 4, 40 items for upper grades). Adding the two scores provides a total score (Total IQ) said to be a measure of general intelligence. All items are pictorial and multiple choice (five choices) and test materials were designed so as to minimize dependence on skills learned in school. For example, in one of the Verbal subtest items the children are asked to mark with a crayon the object, among five objects, that can be eaten; in a Reasoning subtest item, they must cross out the one drawing out of five that does not follow the same rule as do the others. Examples are given in Rosenthal and Jacobson (1968b). Speed of response is not a test variable, but note that the Verbal items are read aloud by the teachers, who must frequently roam the aisles to make certain the children are on the correct page and understand the instructions, whereas the Reasoning items are self-administered.

TOGA raw scores are converted to mental ages (MAs) and tables of MA equivalents are given in the TOGA manual, including MA equivalents extrapolated beyond the range used in the normative sample. Total IQs are derived from the MAs via the ratio formula (MA/CA X 100). However, IQ conversions were not extrapolated below IQ 60 or above IQ 160, and extrapolations of IQs beyond those limits are discouraged. Consequently, the *Pygmalion* IQs that were below 60 and above 160 were based on inappropriate extrapolations. The directions for administering the test give the testers a good deal of freedom, an important consideration in *Pygmalion*, where teachers administered the test.

Note also that group intelligence tests are in general more subject to artifacts than are individual tests, although Rosenthal and Jacobson (1968a) defended their use of a group test on logistical grounds (size of the sample) and suggested also that, compared with individual testing, group testing provided a better safeguard against "unintended effects of the examiner's expectancy" (p. 70). Elsewhere they suggested that "the unreliability of the instrument makes the results the more dramatic [because] ... as test reliability decreases a more robust relationship must exist between the instrument and other variables for these relationships to become significant statistically" (Rosenthal & Jacobson, 1968c, p. 253). In other words, they believed their results were valid because they were found *despite* the reduced reliability of group tests, whereas their critics questioned the results *because* of the test's suspect reliability.

Measuring instruments are often inaccurate and inconsistent. For instance, a blood pressure gauge is, on average, a quite reasonable measure, but there are wide fluctuations over the course of a day, and differences depending on who is administering it. Sometimes an instrument from a particular manufacturer is defective, and some instruments are simply better than others. Intelligence tests are vulnerable to all these problems and more, because training and practice can raise performance. Too often what is mistakenly believed to be a change in intelligence is merely a temporary change in IQ. But IQ tests and intelligence are not the same thing. Sound IQ tests are usually reliable but certainly fallible measuring instruments that infer general intelligence, whereas general intelligence (or *g*) is considered by most intelligence researchers to be, from the time of childhood, a very stubborn individual trait that is the product of genetic–environment interactions (Gottfredson, 1997).[4]

The claim that there are methods to raise intelligence substantially and permanently has a long history, but has produced only repeated disillusion (Spitz, 1986). Consequently many workers become skeptical when rather ordinary events are said to modify intellectual level in more than a trivial way. That Rosenthal and Jacobson were proposing that teachers were raising general intelligence and not simply test scores is evident from their book's subtitle, *Teacher Expectation and Pupils' Intellectual Development*, as well as in their use of such terms as intellectual growth, intellectual development, intellectual gains, and intellectual competence to describe the study and its results (Rosenthal & Jacobson, 1968a,b,c). Furthermore, Rosenthal (1985) continued to use the analogy of children who "could become brighter when expected to by their teachers" just as "rats became brighter when expected to" (p. 44) by their experimenters.

This brief digression was necessary because subsequent criticisms of *Pygmalion* so frequently questioned whether the TOGA should have been administered by the classroom teachers, whether it was an appropriately normed test for children in the lower grades of a somewhat depressed socio-economic area, and whether the gains in test scores reflected raised general intelligence or simply changes in test behavior.

## Back to the Experiment

After the May 1964 testing—that is, during the summer of 1964—20% of the students (who in September would advance one grade, from grades K to 5 to grades 1 to 6) were chosen at random as potential "bloomers." This amounted to an average of about 5 students in each of the 18 classrooms (three classrooms at each grade level), although the number of experimental children in each classroom who were listed as bloomers actually ranged from 1 to 9 because "it was felt to be more plausible if each teacher did not have the same number or percentage of her class" (p. 70). The remaining children served as controls. At the start of the fall semester and four months after the initial May testing a sheet of paper was distributed to each of the 18 teachers listing the names of from 1 to 9 children who would be in the teacher's class that year and who had scored in the top 20% of all Oak School pupils on the Harvard Test of Inflected Acquisition. The teachers were told that they were given the list because they might be interested "to know which of their children were about to bloom" (p. 70), and were cautioned not to talk about the test findings with the children or parents.

In January 1965, at the end of one semester, the students were again administered the disguised TOGA (to test for possible early effects), and again at the end of the school year, in May 1965, 12 months after the pretest and some 8 months after the school year started. This latter (third) test was referred to as the basic post-test because it served as the crucial criterion for gains. At the second and third testings the children took the same test form they had taken at the first test (the pretest) because those who had been pretested in grades K to 1 were, after a year, still within the test level of form K-2, and ditto for the other grades, but any practice effect of repeating the same form of the test should have affected the experimental and control groups equally.

In May 1966, 2 years after the pretest, the children were given the test for the fourth and final time, but by their new teachers who presumably did not know which children had been designated bloomers. This time some of the classes were given the same form a third time, and some even a fourth time, while others classes had advanced to the succeeding test form. Formerly sixth graders were not tested because they no longer attended Oak School. "All tests . . . were scored twice, and independently, by research assistants who did not know which children were part of the control group and which were part of the experimental group" (Rosenthal & Jacobson, 1968a, p. 69).

In sum, the study was designed to measure "whether those children for whom the teachers held especially favorable expectations would show greater intellectual growth than the remaining or control-group children" (p. 68) when tested roughly 5 months (the second test), 8 months (the third test, or basic post-test) and 20 months after experimental treatment (classes) began.

## Results

After 8 months in the classroom the mean gain in Total IQ on the basic post-test for the 65 children designated as bloomers was 12 IQ points, compared with a mean gain of 8 for the 255 control children (I have rounded many of the scores). The differential gain of 4 IQ points was statistically reliable. However, this finding is due almost entirely to combined grades 1 and 2, consisting of six classes with a total of 19 experimental children. In combined grades 1 and 2 the experimental group ($N = 19$) gained an average of 20.5 to the control group's ($N = 95$) average gain of 9.5 in Total IQ. In none of grades 3 to 6, or in combined grades 3–6, were group differences reliable. In fact, in grades 1 and 2 the results were due to five of the six classes because the control group in one of the three second grade classes (with only three experimental children) outgained the experimental group, although not significantly so. Consequently, 16 children in five combined first and second grade classes contributed disproportionately to the significant findings of the total group.[5]

Analyses were also made separately for the Verbal and Reasoning IQs, as well as for boys compared with girls. At the basic post-test there were no statistically significant differences in Verbal IQ gain scores for the entire group or for combined grades 3 to 6. However, combined grades 1 and 2 experimental children outgained control children in mean Verbal IQ by a statistically reliable 10 IQ points, the result of a 14.5-point average gain by the experimental children and a 4.5-point average gain by controls. Again, five of the six classes in grades 1 and 2 were responsible for the significant results because the control group in one of the three second-grade classes (a different class but also with only three experimental children) outgained the experimental group, although not significantly so.

In Reasoning IQ the entire experimental group outgained the control group reliably, but, as in the other analyses, when analyzed separately (combined grades 1 and 2 and combined grades 3–6) only combined grades 1 and 2 produced significant results. The experimental group in combined grades 1 and 2 outgained the control group by 13 points, the result of an astonishing 40-point average gain by the experimental children and an equally astonishing (considering they were the control group) 27-point average gain by the control children. Concerning this remarkable gain by control children, Rosenthal and Jacobson (1968a) conjectured that "it may be that experiments are good for children even when the children are in the untreated control group" (p. 78). They later commented that the results were consistent with the Hawthorne effect: "Perhaps the very fact that university researchers, supported by federal funds, were so interested in the Oak School may have elevated the already good level of morale and teaching techniques shown by the teachers of Oak School" (p. 169).

In the upper grades there were a number of classes in which the control group outgained the experimental group in mean Total and Verbal IQs, although this reversal was significant only for the Verbal IQ gains in one class, a fast track class of third graders with 8 experimental children. (Note also that due to examiner error one of the fifth-grade classes received only the Verbal test and therefore did not contribute to the Reasoning and Total IQs).

Although the between-group gain differences of the individual classes (for *all* comparisons) were rarely statistically significant, the experimental group outgained the control group in Total IQ in 11 of 17 classes, as Rosenthal and Jacobson noted in a footnote (p. 95) and as Rosenthal later reiterated in support of *Pygmalion*. However, statistical support was shaky at best for all but the Reasoning IQ comparisons.

In the comparison between girls and boys on the Reasoning subtest, the experimental girls of combined grades 1 and 2 outgained the control girls by an average of 40 IQ points whereas the *control* boys outgained the experimental boys by an average of 11 points. On the Verbal subtest, on the other hand, the boys and girls contributed roughly the same amount to the approximately 10-point superiority of the experimental over the control group. Consequently the positive results of the expectancy effects on Total IQ of the entire group profited greatly from the exceptionally large gains on the Reasoning subtest by the small sample of experimental girls from five of the six classes (excluding one class where the control group outgained the experimental group) from combined grades 1 and 2.

To provide an empathetic touch, the authors presented a description of 12 of the experimental first and second graders. The changes in scores of the four highest gainers after 8 months in a classroom were indeed remarkable: 133 to 202, 61 to 106, 88 to 128, and 60 to 97. These results were bound to raise some skepticism in anyone inclined to be even a little bit cautious. After all, two children who on pretest had scored in the mentally retarded range (based on their IQs) were, after only 8 months, scoring in the normal range; another, who had been below average, advanced to the superior range; and the fourth went from superior intelligence to the extremely gifted, one might even say the genius, category.

Results of the other two post-tests (other, that is, than the basic post-test) were not considered critical and were left until after the analysis of the basic post-test results. The second testing, one term prior to the basic post-test, was included to gauge how quickly expectancy effects, if any, were produced. The final testing, 1 year after the basic post-test, measured the durability of any effects that were found. When the entire experimental and control groups' mean Total IQ gains from pretest to second testing or from pretest to the final (fourth) testing were compared, neither produced a reliable between-group difference. The only grade that showed a reliable expectancy effect on the final test was (former) grade 5, which had not previously produced an effect. "Why fifth graders, expected to bloom in one year, should show such large expectancy advantages during a subsequent year in a classroom taught by a teacher given no special expectation for their intellectual performance remains a baffling question" (Rosenthal & Jacobson, 1968a, p. 130).

Considering subsequent criticisms that the TOGA was unreliable for the yougest children at pretest, the basic post-test (that is, the third administration of the TOGA) might serve as a more credible baseline for comparison with the final test. For the lowest two grades combined, 95 control children had a mean Total IQ at the basic post-test of 101, which at the final testing (for the 75 control children still available) remained unchanged. Nineteen experimental children had a mean basic post-test Total IQ of 117, which at the final testing (for the 15 remaining experimental children) dropped to 109. Moreover, "When only those children [from the entire group] were considered who had taken the one year post-test [the basic post-test] and the two-year follow up test there was a significant reduction of expectancy advantage in total IQ ($p < 0.05$, two-tail)" (Rosenthal & Jacobson, 1968a, p. 129). This comparison varied from the authors' usual practice of using the scores of all participants at each post-test for their major analyses, choosing not to discard the earlier scores of subjects who had dropped out along the way. But the results were no different, confirming that if there was any true initial effect it was ephemeral.

Perhaps the strangest aspect of this study was the finding that most teachers paid no attention to the names of the children expected to spurt academically. In June of 1966, a few weeks after the last administration of the TOGA and roughly a year after the end of the experimental period, 16 of the original 18 teachers were still available. After being told at a group meeting about the nature of the experiment, each was interviewed individually. The authors described the teachers' reactions as "startling."

> While all teachers recalled glancing at their lists, most felt they paid little attention to them. Many teachers threw their lists away after glancing at them. Many of the teachers felt there were so many memos coming from the office that first week of school that the list of names was just another list and got no special attention. (Rosenthal & Jacobson, 1968a, p. 154)

The teachers were unable to recall the names of those children in their classes of the previous year who had been designated as potential academic spurters, and had trouble even recognizing them.[6] Of the 72 children originally in the experimental group, the teachers correctly recalled 18 experimental children, but incorrectly recalled 18 control children, as being on their list. There was no difference when the gain scores of the experimental children that teachers correctly recalled or recognized were compared with the gain scores of those they did not. In fact, none of the second-grade teachers recalled correctly any of the children designated as potential bloomers, yet this grade had the second largest differential gain. Of course, one can always argue that the names registered unconsciously. Rosenthal and Jacobson (1968c), however, raised the possibility that, following the laws of forgetting, "whatever mediated the effects of teachers' expectations operated early in the academic year" (p. 246).

In their *Scientific American* article, Rosenthal and Jacobson (1968b) compared the gains made in their study, when "the only people affected directly were the teachers" (p. 23), with the more modest gains made by the federal Elementary and Secondary Education Act, which focused on the child. In view of their results, they wrote, perhaps "more attention in educational research should be focused on the teacher" (p. 23). They recommended teacher expectancy as an efficient and economical method for attacking the problem of low intelligence and poor educational performance. This did not sit too well with teachers (e.g., Shanker, 1971), who now had the additional responsibility of raising children's intelligence.

## A Similar Study Mentioned in *Pygmalion*

In their book, Rosenthal and Jacobson described a dissertation by Flowers (1966) that, although designed independently, greatly resembled *Pygmalion*. It was an understandable concordance because in his related literature Flowers integrated Rosenthal's studies on the self-fulfilling prophecy with the work of other social psychologists who "ascribed the low academic achievement of culturally disadvantaged children to the poor performance teachers expected" (p. 64). He designed his study to measure "whether or not there would be an observable difference in the achievement of disadvantaged students after they had been taught by teachers who were led to believe they had higher tested achievement or ability" (p. 1). In addition to academic achievement, he measured changes in scores on a group intelligence test (the Otis Quick-Scoring Test of Mental Ability). His subjects were children in the seventh grade of two widely separated junior high schools, one in the midwest and one in the east, both of which were in depressed areas. Because these schools used an ability-track system (as did the Oak school) it was possible to match control and experimental children on scores they had achieved in sixth grade and then shift the experimental children (all of whose scores warranted placement in the average classroom) to the top classroom while their controls remained in the average classroom. This set up a natural teacher expectancy for higher performance by the experimental children.

A year later, in the eastern school the experimental group lost a mean of 1 IQ point while the control group remained at the same level. In the midwestern school the experimental children gained 2.16 IQ points while the control group (unlike *Pygmalion*) dropped 2.68 IQ points, a reliable difference. Because there were no significant differences in reading and arithmetic, and because of his split results, Flowers concluded that his hypothesis was not supported.

Rosenthal and Jacobson (1968a) suggested four possible reason that "the gains demonstrated were not dramatic," one of which was, "Even the nondramatic gains demonstrated in Flowers' study may have been too high and, in fact, educational self-fulfilling prophecies do not occur" (p. 59). Rosenthal never again made so striking a statement, even couched as a possibility.

## Immediate Critical Reaction to *Pygmalion*

Although there were many favorable reviews of *Pygmalion in the Classroom*, of the reviews that were unfavorable two in particular were extremely critical. The first, by Thorndike (1968), contained his prescient second sentence: "In spite of anything I can say, I am sure it [Pygmalion] will become a classic—widely referred to and rarely examined critically," followed by his unusually blunt lament, "Alas, it is so defective technically that one can only regret that it ever got beyond the eyes of the original investigators!" (p. 708). For Thorndike, the data indicated that something was wrong with the TOGA and/or the testing procedure. Regarding the mean Reasoning IQ of 31 for one classroom of 19 children about to enter first grade at the time of the pretest, he remarked that the children "just barely appear to make the grade as imbeciles!" (p. 709), and he concluded that these kinds of data "show that the testing was utterly worthless and meaningless" (p. 710). In fact all 63 children entering first grade had a mean Reasoning IQ of 58. To achieve so low a score Thorndike estimated that they needed a raw score of only about 2, which is below the score one would obtain by chance. On the other hand, by his calculations the post-test Reasoning IQ of 150 for the six "spurters" in the fast track classroom of the second grade, assuming they had a mean chonological age (CA) of 7.5 years, would have required perfect scores, so he wondered how they could have had an S.D. of 40.17 (actually 40.71).

In his response, Rosenthal (1969a) noted that once Thorndike had questioned the validity of the TOGA's Total IQ measure for the lower grades he never returned to the Total IQ "apparently because it was too well measured" (p. 689), a presumptuous statement that was startling in view of the Total IQs' chaotic fluctuations. Rosenthal then commented that Reasoning IQ gains for the experimental children were greater than for control children in 15 of 17 classrooms. To explain the IQ of 150 obtained by the six children in the fast track classroom of second grade, he pointed out that "their mean MA [mental age] was simply 1.5 times the magnitude of their mean CA. The MAs were 16.5, 16.5, 10, 10, 10, and 8.9" (p. 690). In other words, using the formula IQ of MA/CA × 100, an IQ of 150 results when the MA is 1.5 times the CA.

But this indicates that there must have been a large age disparity within the class. Two of the six children must have been roughly 11 years of age to have obtained an MA of 16.5 and an IQ of 150 (16.5/11 × 100 = 150), while another child must have been roughly 5.9 to 6 years of age to have obtained an MA of 8.9 and an IQ of 150 (8.9/5.9 × 100 = 150).

Concerning the astonishingly low IQs of the first grade children when they had been pretested in kindergarten, Rosenthal explained, "These low IQs were earned because very few items were attempted by many of the children" (p. 690). Furthermore, he continued, the scores on the TOGA Reasoning subtest predicted, at an above chance level, the ability track that kindergarten teachers recommended for the children entering first grade. Likewise, a year later there was a significant correlation of 0.49 between the Reasoning IQ pretest and the first grade teachers' assessments of future success of their students.

The rejoinder by Thorndike (1969) was brief. He agreed that there is a table to convert total raw scores on the TOGA to age equivalents (MAs) ranging from 0.5 to 16.5 years, a table obviously used by Rosenthal. But, he went on, age equivalents are "about as unsatisfactory an approach to an equal-unit scale as we have," particularly when "extrapolated far beyond the ages or grades in which testing [standardization] was done" (p. 692). It is nonsense, he wrote, to assign an MA to a child who "did not understand what he was supposed to do and consequently omitted all or most of the items" (p. 692). Furthermore, that would allow room for teachers to influence scores on the post-tests by encouraging pupils to guess at a few more items whether or not they knew the answers. "Normal luck could then produce a measurable if not substantial increment in average score" (p. 692).

The second very critical review was by Snow (1969), who had sent for and received the raw scores from Rosenthal. Snow's review and the subsequent analyses by him and his colleagues have been by far the most extensive critiques and, as noted, the latest skirmish between Snow and Rosenthal was as recent as 1995. Like Thorndike, Snow was unsparingly pejorative. "*Pygmalion* inadequately and prematurely reported in book and magazine form, has performed a disservice to teachers and schools, to users and developers of mental tests, and perhaps worst of all, to parents and children whose newly gained expectations may not prove so self-fulfilling" (p. 199).

Snow reiterated questions about the TOGA, which "does not have adequate norms for the youngest children, especially for children from lower socio-economic backgrounds" (p. 198), and, as did Thorndike, he mentioned improbable pretest mean Reasoning IQs of 31, 47, and 54 for some of the first grade classes. Drawing from the original data, he supplied the reader with additional scores of individual participants: one child with a pretest Reasoning IQ of 17 and post-test IQs 148, 110, and 112, another with a pretest Reasoning IQ of 18 and post-test IQs of 44, 122, and 98, and a third with successive Verbal IQs of 183, 166, 221, and 168. Seven of these 12 scores were beyond the TOGA's norm range of 60 to 160, and therefore were extrapolations. Snow described some of the "serious measurement problems and inadequate data analysis" (p. 197) in some detail and promised a full report with a further reanalysis of the data.[7]

Jensen (1969, pp. 107–108) briefly discussed *Pygmalion* in his well-known *Harvard Educational Review* article on the failure of compensatory education to permanently boost IQs. He too questioned the dubious practice of having teachers administer the group test. There would have been no statistically significant results, he asserted, had the means of the group comparisons for each classroom been used rather than the scores of individual children. In reply, Rosenthal and Rubin (1971) accused Jensen of "sweeping-under-the-rug ... undesirably low *p* values" (p. 144), and later Rosenthal (1973) added that he and Jacobson had in fact compared classrooms, resulting in even larger effects. Jensen's objection that teachers should not have administered the test was groundless, he argued, because when the children were retested by people who knew nothing about the experiment the expectancy effects actually increased.

## Rosenthal's Attempts to Replicate and Extend *Pygmalion*

There were a number of studies through 1971 that were impelled by *Pygmalion*, including three in which Rosenthal participated. In a footnote on p. 96, referring to a speculation on why boys and girls differed in gains made on Verbal compared with Reasoning IQs, Rosenthal and Jacobson (1968a) reflected

> on the complexity of nature and the need for noncomplacency in the behavioral researcher. Preliminary results of a study conducted with Judy Evans give just the opposite results [as *Pygmalion*] and with an equally significant probability level. The same basic experiment conducted at Oak School was repeated in two elementary schools located in a small Midwestern town . . . No expectancy advantage was found for either boys or girls as measured by either total IQ or verbal IQ . . . But now we know for sure that Oak School's results, like the results of all behavioral experiments, are not universal (p. 96).

There was no advantage in Reasoning IQ either, because the experimental boys gained 8 points more than the control boys, whereas the experimental girls lost 10 points more than the control girls.

Rosenthal and Jacobson (1968a) pointed out that the midwestern children of Evans and Rosenthal (1969) were drawn from a middle-class community whereas the western Oak School children were from a lower-class community and "included a large proportion of minority group members" (p. 96). One finding did replicate *Pygmalion*, however. At the year-end interviews the "teachers were remarkably inaccurate in their memory of these [experimental] children's names" (Evans & Rosenthal, 1969, p. 371). In describing this study in his chapter in *Artifact in Behavioral Research* (Rosenthal & Rosnow, 1969), Rosenthal (1969b) commented that, "Just as in the West Coast experiment [Pygmalion], however, all the children [both experimental and control] showed substantial gains in IQ. These results, while they suggest the potentially powerful effects of teacher expectations also indicate the probable complexity of these effects as a function of pupils' sex, social class, and, as time will no doubt show, other variables as well" (p. 263). The suggestion here that the "potentially powerful effects" of teacher expectation raised the intelligence of both groups ignores the even more potentially powerful role of practice effects.

Rosenthal was the second author of a "preliminary report" during this period (Anderson & Rosenthal, 1968). Twenty-eight boys from a state school for people with mental retardation were considered ideal subjects because "expectancies held by staff, teachers, and attendants about the retarded child tend to become self-fulfilling prophecies" (p. 479).[8] For the first condition, 25 counselors in a summer day camp were told which 11 boys the Test of Inflected Acquisition (actually the TOGA) predicted "were likely to show unusual intellectual growth" (p. 479). Additionally, 6 of these 11 experimental boys and 9 of the 17 control boys were randomly chosen for a second condition, a one-to-one tutoring relationship with 15 volunteer high school students 2 or 3 h two nights a week during the 8 weeks of camp, allegedly to get to know the boy and help him learn to read.

After 8 weeks, 25 of the boys (with an average IQ of 46) were still available for retesting on the TOGA. Results indicated no significant effect of expectancy on the Total IQs. The only significant expectancy effect was on the Reasoning IQ of the TOGA for the six boys who were expected by their counselors to bloom intellectually and who also received tutoring. They showed an expectancy *disadvantage* of nearly 12 points; that is, their scores *decreased*.

Rosenthal (1969b) described follow-up testing 7 months after the 8-week summer session. The boys who had been in both the expectancy and tutoring condition made up their "expectancy disadvantage" in Reasoning IQ and were scoring at the same level as the control group, both groups now showing a 4-point loss. But inexplicably, the boys who had been given either tutoring or favorable counselor expectations 7 months earlier now showed significantly greater advantage in Reasoning IQ compared to boys given both or neither. Considering the advantages of each of these conditions separately, why did not boys show an expectancy advantage when these conditions were combined? Rosenthal suggested one possible explanation: "the simultaneous presence of both treatments [favorable expectancy plus tutoring] led the boys to perceive too much pressure" (Rosenthal, 1969b, p. 264).

Despite the extremely small samples and confusing results, this study was included in the chapter in *Artifact in behavioral research* by Rosenthal (1969b) as suggesting that "teacher expectations can significantly affect students' intellectual performance in a period as short as two months" (p. 263).

A study by Conn, Edwards, Rosenthal, and Crowne (1968) was described in *Pygmalion* as a quasireplication carried out in a "middle- or upper-middle-class community some 3000 miles from Oak School" (p. 140) (obviously on the east coast). The procedure was the same as in *Pygmalion* except that the TOGA pretest was given to grades 1 through 6 at the beginning of the second semester rather than the first, which meant that the teachers had these children in their classes for one semester before being provided with the list of potential bloomers. The teachers were advised that "scores on the 'test for intellectual blooming' indicated they [the listed children] would show unusual intellectual gains during the next year" (p. 28). After one semester in the treatment condition there were no statistically significant effects, a result, according to Rosenthal and Jacobson (1968a), that was not unexpected because the teachers already had a full semester to form their own expectations. Group differences were of borderline significance ($p = 0.08$) a year after the experiment ended, but they had *reversed*: the control children gained 4.58 more in IQ than the experimental children. Rosenthal and Jacobson speculated that perhaps "they suffered a relative deprivation in moving into a classroom in which the teacher had no special expectation for their intellectual growth and this disappointment may have been reflected in their intellectual performance" (p. 145). Concerning their results in general, Conn et al. (1968) concluded that "positive expectations do not necessarily lead to positive results for all pupils. The process is one of much greater complexity, involving both situational factors and the perceptions and other characteristics of the individual student" (p. 33).

## Relevant Dissertations Through 1971

From 1968 to 1971, *Pygmalion* was instrumental in generating a cascade of doctoral dissertations on teacher expectancy effects. The essential elements of those that included intelligence test performance are briefly reviewed here and, along with all the studies reviewed here, are summarized in Table 1.

Based on his 1968 doctoral dissertation, the published article by Claiborn (1969) described his attempts to avoid such *Pygmalion* artifacts as pre- to post-test difference scores "not corrected for known pretest differences, and partially attributable to regression effects" (Claiborn, 1969, p. 378). As with the three replications in which Rosenthal participated, this was best described as a partial replication. By using observers (raters) in the classroom, he tried to determine if there were changes in the teacher–pupil interaction of those teachers given the (false) information about pupils expected to blossom

**Table 1.** Studies of the *Pygmalion* Effect on IQ

| *Study* | *Sample* | *Raising Expectancy* | *Duration/IQ Test* | *Results* | *Comments* |
|---|---|---|---|---|---|
| Flowers, 1966 (Dissertation) | 7th grade disadvantaged children, one class in east, one in midwest | Exp. children placed in higher track than warranted | School year/ Otis | In 1 of 2 classes, Exp. children up 2 pts, Controls down 3 pts. ($p < 0.05$) | In other class, no difference |
| Rosenthal and Jacobson, 1966, 1968a | Grades 1–6, lower class, in Southern California 17% Mexican | Misled teachers on Exp. children's potential | 8 months. Post-tests given up to 2 years/ TOGA (disguised) | Exp. children gained 12 pts. (4 pts. over Controls) due to grades 1–2 | Wildly fluctuating IQs in repeated tests |
| Anderson and Rosenthal, 1968 | Boys with mental retardation, 9–16 yrs.old | Misled Counselors. Half the Exps. *S*s ($N = 6$) were also tutored | 8 weeks/TOGA (disguised) | No difference in Total IQ between Exps. and Controls (no effect) | The Exp. *S*s who were also tutored lost 10 pts. in Reasoning IQ |
| Conn et al., 1968 | Grades 1–6, upper-middle-class in east | Same as Rosenthal and Jacobson, 1968a | 4 months (2nd semester)/ TOGA (disguised) | No statistically significant differences | Differences a yr. later in unpredicted direction |
| Evans and Rosenthal, 1969 | Grades 1–6, middle-class in midwest | Same as Rosenthal and Jacobson, 1968a | School year/ TOGA | No difference in Total IQ between Exps. and Controls | Sex differences in Reasoning IQ results |
| Claiborn, 1969 (based on Dissertation) | 1st grade classes, middle-class children | Same as Rosenthal and Jacobson, 1968a | 2 months (2nd semester)/ TOGA (disguised) | No significant effect (but *all* pupils' mean IQ gain = 11 pts.) | Raters in half the classrooms periodically |
| José, 1969 (Dissertation) | 1st and 2nd grades, 7 schools, varied backgrounds | Same as Rosenthal and Jacobson, 1968a | 16 weeks/ TOGA (disguised) | No significant effect | Observers visited classrooms periodically |
| Kester, 1969 (Dissertation) | 7th grade classes, 6 schools, middle-class | Misled teachers on experimental children's IQs | 9 weeks/ Otis–Lennon | No significant effect (see my Footnote 9) | Observers visited classrooms periodically |
| Carter, 1970 (Dissertation) | 7th graders in one junior high, range of *S*s, partitioned into two groups | Misled teachers on experimental children's IQs | About 7 1/2 months/ Lorge–Thorndike, Verbal, Level 4 | Control–Exp (all *S*s) pre- to post-test difference = 2.82 ($t = 1.78$, n.s.) | Unusual No. of decreases in IQ resulted in difference in 1 group ($p < 0.05$) |
| Maxwell, 1970 (Dissertation) | 2nd and 4th grades in a parochial, likely middle-class, school | Misled teachers on experimental children's IQs | 7 months/ Stanford–Binet | Exp. *S*s' 5.59-pt. rise significantly greater than Controls | Unique unqualified support |

(continued)

**Table 1.** Continued

| *Study* | *Sample* | *Raising Expectancy* | *Duration/IQ Test* | *Results* | *Comments* |
|---|---|---|---|---|---|
| Keshock, 1970 (Dissertation) | 2nd–5th grade disadvantaged African–Americans | Misled teachers on experimental children's IQs | 9 months/ Stanford–Binet | Essentially no change in mean IQ for Exp. or Control groups | One of three studies from Case Western Reserve |
| Ginsburg, 1970 (Dissertation) | 5 lower- and 5 middle- and upper-income 1st grades | Misled teachers on experimental children's IQs | School year/ TOGA | No significant effect | All groups, gained in IQ |
| Grieger, 1970 (Dissertation) | Rural 1st–4th grades, lower middle–class | Same as Rosenthal and Jacobson, 1968a | 2 mos./Cal. Test of Mental Maturity | No significant effect | Periodic observers. All groups gained in IQ |
| Henrikson, 1970 | Disadvantaged Head Start *S*s now in kindergarten | Misled teachers that Exp. children in top quartile | School year/ Slosson | No significant effect | 53% African–American, 25% Spanish Surnames |
| Fleming and Anttonen, 1971b | 2nd grades from 22 schools, lower and middle-class | Misled teachers on experimental children's IQs | School year/ Kuhlmann–Anderson | No significant effect. All groups gained in IQ | Of 1087 *S*s recruited, 895 remained for entire study |
| Fielder, Cohen, and Feeney, 1971 | Grades 1–6, 36 classes, 24 with many Mexican Americans | Same as Rosenthal and Jacobson, 1968a | 4 months in 2nd semester/ disguised TOGA | No significant effect | Commented on Difficulty giving TOGA to 1st graders |
| Fine, 1972 (Dissertation) | 2nd graders from 5 schools, low socioeconomic | Misled teacher on Exp. *S*'s reading potential | 2nd semester/ Cognitive Abilities Test | No significant effect on IQ (or reading). Both groups raise IQ | About 2/3 African–American, 1/3 Caucasian |
| Pellegrini and Hicks, 1972 | Elementary school, lower class, 70%–80% Mex–American | Misled part-time tutors on Exp. children's IQs | 17 weeks/ Peabody + Simil. subtest of Wechsler | Sign. gains only when tutors were briefed on IQ test material | Suggests "teaching to the test" in some studies |
| Rosenthal, Baratz and Hall, 1974 | Grades 1–6, 96% lower class African–American | Misled teachers on *S*'s "creative potential." | School year/ disguised TOGA | No significant effect on Total IQ for combined group | Exp. *S*s in 5th grade outgained Controls |
| Sutherland and Goldschmid, 1974 | Grades 1–2 in 3 schools, middle-class | Compared teacher rankings with IQ change | 5 Mos./4 subtests of Wechsler + Lorge–Thorndike | No effect on IQ gain of higher (but wrong) ranking | An effect of lower (but wrong) ranking |

intellectually. Expectancies were introduced roughly 1 month into the second (spring) term, and final retests were administered 2 months later. This was a short experimental period, but Rosenthal had also participated in replications using relatively short experimental durations (Anderson & Rosenthal, 1968; Conn et al. 1968).

At the end of the experiment the teachers' responses on a questionnaire indicated that they accurately remembered the potential bloomers. Nevertheless, Claiborn found no expectancy effects on TOGA Total IQ, and great variability depending on schools, classes and conditions. Nor did raising teacher expectancies affect teacher–pupil interactions. Claiborn raised the possibility that because the teachers must have formed an impression of the students prior to the study (which started a month after the second term began), teacher expectancy effects on intelligence, if they exist at all, are ineffective in overcoming an established impression (see also Conn et al. 1968).

Referring to Claiborn's study, Rosenthal and Rubin (1971) pointed out that "two of the three teachers whose experimental condition was similar to that of the RJ [Rosenthal and Jacobson] study were either fully aware or partially aware of the nature and purpose of the experiment" (p. 150), a point mentioned in Claiborn's dissertation but not reported in his journal article. They were referring to Claiborn's evaluation of teacher awareness, a potential problem because observers were in the classroom part of the time. On the basis of their responses to a questionnaire, eight of the 12 teachers were classified as unaware, two as moderately aware and two as fully aware. Separate analyses on the classes in "the two schools in which the teachers showed no awareness of the nature of the experiment" (Claiborn, 1968, p. 60) revealed no major significant expectancy effects on Total IQ.

The dissertation of José (1969)—the basis of the subsequent article by José and Cody (1971) (although unmentioned by them)—also included teacher behavior as one of the variables. At the start of the spring term the teachers were told that the observers were studying teacher–pupil interaction for a different study than the parallel study of identifying late-blooming students. A preliminary measure of teacher–student interaction was obtained, the disguised TOGA was administered (not by the teachers but by "naive" assistants, and scored with students' names covered) and each teacher was given the names of the four students who, according to the test, were academic bloomers. Copies of the Rosenthal–Jacobson printed explanation of the "Study of Inflected Acquisition" (with minor changes) were also distributed. The observers, uninformed as to the identity of experimental and control children, visited the classrooms at the end of the first week and every 4 weeks thereafter for the 16 weeks of the project.

Results revealed no reliable differences in the changes from mean pre- to mean post-test TOGA Total IQ of the experimental compared with the control children, nor were there any reliable differences in teacher behavior toward the two groups. At the end of the study 11 of the 18 teachers stated on a questionnaire that they had not expected more from the potential academic bloomers (according to some, because they knew the children and their background and therefore knew the child's potential). However, no reliable IQ effects were found for the seven teachers who expected their "potential bloomers" to bloom, nor did reliable effects emerge even for the experimental students of four of these seven teachers who believed their special children really did improve.

As with José's, the dissertation of Kester (1969) was unacknowledged in the spin-off article by Kester and Letchworth (1972). Kester measured the effects of teacher expectancy on many variables and almost as an afterthought included the intelligence test scores (change in IQ was not one of his 10 hypotheses). Teachers were told that a few above-average children had been placed in their average sections, and that those students, as well as the other students in the class, would be tested the first 2 days of school, and again in 9 weeks (the intelligence test was the group-administered Otis–Lennon Mental

Abilities Test). At no time, Kester warned, were the bright children to be made aware that they are the experimental subjects.

After the pretests and the assignment of subjects, teachers were provided a list of their students who were (allegedly) exceptionally bright, with an IQ higher than 120. Teachers were also informed that for purposes of the experiment someone would observe the classes four or five times over the course of the 9-week period. In actuality, all students who scored between 90 and 110 on the Otis were randomly assigned to the experimental or control group until there were 75 experimental students (falsely labeled superior) and 75 control students scattered through the seventh grade English and math classes.

No significant differences in the mean IQ change of the experimental compared with the control group were found, despite the fact that the teachers communicated more positively and for longer periods of time with the allegedly superior students than with the rest of the students. However, the English and mathematics teachers comprised only one-third of the students' teachers.[9]

In the dissertation of Carter (1970), student records were altered so that, among other changes, experimental students were purported to have IQs 7 to 15 points higher than their true IQs. Each member of matched pairs was randomly assigned to either an experimental or control group. Twelve pairs who, based on their sixth grade performance, had been ranked as higher level students, were in Section A (rounded true mean IQs = 120 and 122 for the 12 experimental and their paired control subjects, respectively). Ten pairs, who had been ranked in the lower level, were in Section B (true mean IQs = 108 and 110, respectively). The average upward adjustment for experimental students in Section A was 9 points and in Section B was 12 points. The altered records were given to all seventh grade teachers but only two teachers—a male English teacher and a female social studies teacher—were chosen to participate because of their unfamiliarity with the students' backgrounds.

Our interest is in the IQ changes from pretest during the first week in October 1968 to post-test during the third week in May 1969, an interval of more than 7 months. Testing was by the school counselor in class sessions. For the full comparison between the 22 experimental students and their 22 paired controls, the experimental group decreased 1.41 IQ points and the controls decreased 4.23 points ($t = 1.78$), a group difference reported as significant using a one-tailed test. However, it is an unreliable difference using a two-tailed test. When analyzed separately, the subgroup of 10 experimental students in Section B (the students of average intelligence) gained 2 points and their controls decreased by 3.5 points, a statistically significant group difference, two-tailed. In Section A (the students of above average intelligence, where there was less room for gains), the subgroup of 12 experimental students and their controls both decreased 4 to 5 points. Note that of the total of 44 students only 12 (8 experimental, 4 control) increased their IQ, 4 (2 of each) showed no change, and 28 (12 experimental, 16 control) decreased. The most striking aspect of this study is the unusual number of decreases in IQ scores, perhaps in part reflecting regression to the mean in so many above average students.

Two dissertations, those by Maxwell (1970) and Keshock (1970), and one study (Fleming & Anttonen, 1971b, discussed later), issued from Case Western Reserve University. All three used essentially the same procedure but with different populations. One of them, that of Maxwell (1970), was the only dissertation of all those reviewed here that provided unqualified support for *Pygmalion*. Among other things, it did not depend on control group losses to produce an effect. It differed from *Pygmalion* in a number of ways, including the fact that it did not mention academic or intellectual blooming. Along with other tests, Maxwell administered the individual 1960 Stanford–Binet Intelligence Test to 64 students who, after the summer, would enter the second and fourth grades of their parochial school located in what he called a "bedroom community." The average pretest IQ was about 109. Test results were given to a graduate student to randomly assign 32 students to the experimental group (whose IQs were then raised 16 points) and 32 to the control group (whose IQs were left unchanged). These IQs and their percentiles were delivered in sealed envelopes to the school principal who, about 1 week after school started in September 1969, opened and distributed them to each teacher, at which time the score of each student was discussed. The principal knew about the study, but not which scores were falsely inflated. The examiner (Maxwell) was not informed about the students' placements so as not to influence his administration of the tests.

Seven months after the reports were given to the teachers the tests were re-administered by Maxwell, still unaware of the children's status. The mean IQ of the experimental group rose from 108.57 to 113.16, whereas the control group remained unchanged at 110.32, a statistically reliable difference. Of incidental interest, there was no reliable effect of grade level or sex, and the control group showed no Hawthorne effect.

Participants in the dissertation of Keshock (1970) were 48 second through fifth grade disadvantaged African–American boys in an inner city private school. In August 1969 the participants were randomly chosen, with the constraint that they have IQs between 84 and 115. At pretest a number of different tests, including the 1960 Stanford–Binet Intelligence Test, was administered by Keshock himself. The procedure was essentially the same as in Maxwell (1970) who, incidentally, had the same doctoral committee chairman. After roughly 9 months Keshock (still uninformed as to the status of the children) administered the post-tests, including the Stanford–Binet. Over the 9-month period there was essentially no change in mean IQ for either group.

Ginsburg (1970) also used the popular method of inflating IQs to raise teacher expectancy. The study took place in one first grade class in each of 10 schools, 5 that were receiving Title I funds as underprivileged areas—and for whom teachers presumably would have low expectations—and 5 middle- and upper-income schools in the same school district. During the second week of September 1969 each of the 10 teachers rated each child on scales of intellectual achievement and academic functioning. The following week Ginsburg administered the TOGA. The pretest IQs of 69 of the children was withheld from their teachers, the true IQs of 67 of the children was reported to their teachers, and the pretest IQs of the remaining 65 children were inflated by 10 points before being reported to their teachers. The test information was given to the teachers on a sheet listing the children's test results along with the teacher's initial rating (giving them a chance to compare their ratings with the scores), but the true purpose of the project was withheld and the lack of test scores for the group E-1 children blamed on clerical error or illegibility.

At a post-test in April 1970, all three groups gained in mean IQ, but not differentially. Nor did the teachers' ratings have any effect on the post-test IQs. Both Title I and non-Title I teachers tended to overestimate their students' IQs.

Grieger (1970) followed the methodology of *Pygmalion* more closely by misrepresenting the group-administered California Short Form Test of Mental Maturity as a test being validated to predict which children will show an academic and intellectual "spurt" in the near future. He administered the test to the 18 classes (later reduced to 17) and introduced observers into the classrooms as school psychologists in training who needed to observe what transpires in the classroom. Approximately 20% of the students in each class (3 to 6 students, for a total of 72) were chosen randomly and lists of these alleged potential bloomers were distributed to the teachers. An equal number of children served as controls. About 2 months later, the observers repeated their classroom observations, followed by post-testing. Results were uniformly negative; both groups gained an average of 4 to 5 IQ points and the teachers did not behave more positively toward experimental than control children. At the study's conclusion all teachers were asked to list from memory their intellectual bloomers. Their average recall was 86%.

Henrikson (1970) started with a sample of 76 4- to 5-year-old disadvantaged Headstart children who would be eligible for kindergarten the coming school year. They were pretested by trained aides on the individual Slosson Intelligence Test (along with an achievement test) in August of 1969, just before they were to enter kindergarten. Because of attrition, results were based on 51 of the children. Fifty-three percent were African–American, 25% had Spanish surnames and the rest were non-Hispanic whites. During the first month of school the children were ranked by pretest score, then a child was assigned randomly to either the experimental or control group and the child with the next ranking was assigned to the other group, and so on, thereby assuring ability balance between the two groups. Nineteen children served as experimental subjects, distributed so that there were from 1 to 3 in each of 10 kindergarten classes.

During the second week of school, a letter was sent to each of the 10 participating kindergarten teachers. It described the study as an attempt to assess the educational value and effect on the children of the local Headstart, and listed the names of the children who (allegedly) had been in the top quartile (Henrikson says he should have written "quarter") on tests given at the conclusion of the Headstart program. Teachers' knowledge of the list was verified three weeks later when all 10 correctly listed the "top quartile" children in their classrooms. At the end of the school year there was no significant difference in the mean gain in Slosson raw scores of experimental and control groups.

In 7 of these 10 1966–1970 dissertations (Flowers, 1966; Kester 1969; Carter, 1970; Ginsburg, 1970; Henrikson, 1970; Keshock, 1970; Maxwell, 1970) attempts were made to raise teacher expectancies by either the deceptive enhancement of students' scores or by elevated class placement. Rosenthal's strategy of informing the teachers to expect academic or intellectual blooming, or some variant of it, was used only by Claiborn (1968), José (1969) and Grieger (1970), none of whom replicated *Pygmalion*'s results.

## Additional Studies and Controversy: 1971–1974

Fleming and Anttonen (1971a) recruited a large sample of 1087 second-grade children (of whom 859 remained for the entire experiment) in 39 classrooms from 22 schools for a study of teacher expectancy effects on a number of variables. The variable of interest to us—the effects of teacher expectancy on IQ—was published separately (Fleming & Anttonen, 1971b). At the start of the study, which extended from September 1968 to June 1969, each teacher was randomly assigned to one of four groups, three of which are of interest to us: one to whom the true Kuhlmann–Anderson IQs were reported, another to whom no IQ information on the students was given, and a third in which IQs were inflated by 16 points. The intelligence test was readministered near the end of the school year, in May 1969. All groups gained 6 to 7 IQ points, from about 105 to 106 at pretest to 112 to 113 at post-test. The teachers turned out to be very perceptive. Those who received the inflated scores believed, more so than did the teachers who received true scores, that the test information was inaccurate.

Along with the dissertations of Maxwell (1970) and Keshock (1970), this study completed the trilogy of studies from Case Western Reserve University in which the IQs of experimental students were surreptitiously raised one standard deviation.[10]

Fielder et al., (1971) consulted with Rosenthal before selecting from each of three elementary schools two classes at each of grades 1 to 6 (total of 36 classes). The schools were in Southern California and two of them had many Mexican–American students. Teachers were told about the technique being developed to predict intellectual growth, or "late blooming." The test was the disguised TOGA, administered two weeks before the spring semester and re-administered 4 months later. Other aspects of the study duplicated *Pygmalion* very closely. There were no effects of expectancy on any measure for the combined six grades or for any subgroup. As with other studies, the possibility was entertained that teachers had already formed expectations.

An interesting aspect of this study was the description of the classroom behavior during the 45 minutes it took to administer the TOGA to first grade children. In most of their first grade classes one person read the test while another walked around to answer questions and help the students, and in several classes a third person assisted. Despite the presence of two and sometimes three adults for 23 to 26 students, keeping the children in their seats or preventing them from copying or cheating was "next to impossible . . . We wonder how Rosenthal and Jackson solved this 'activity' problem for first grade *S*'s, especially since their pretest took place when the 'first graders' were still in kindergarten" (p. 1227).

## The Controversy Intensifies: The Elashoff and Snow Book

*Pygmalion Reconsidered* by Elashoff and Snow (1971) included the expanded critique promised by Snow, along with reprints of a number of reviews, a chapter by Rosenthal and Rubin replying to the critique and reaffirming *Pygmalion*, and a rejoinder by Elashoff and Snow. Also included was Baker and Crist's review of the literature on teacher expectancies, in which studies directly attempting to replicate *Pygmalion* were reviewed separately. Doing so allowed them to disentangle the effects of teacher expectancy on intelligence from the effects of teacher expectancy on other kinds of behavior. "Teacher expectancy," they concluded, "probably does not affect pupil IQ . . . [but] may affect pupil achievement" (p. 61).

In their chapters, Elashoff and Snow pointed to what they considered many incorrect and contradictory conclusions and interpretations in *Pygmalion*. They maintained that the summation of results misleadingly suggested that all children gained, whereas in fact the results were for average gains only. As an example of inadequate presentation of data, they noted that the reader could not determine the wide range of IQ scores (30 to 262 for Total IQ, 0 to 262 for Reasoning IQ, and 46 to 300 for Verbal IQ). They mentioned that tables and graphs were presented so as to exaggerate a desired effect. Although the combined first and second grade experimental and control groups differed at pretest—the weighted mean Total IQ of the experimental group was 95.91 compared to a weighted mean of 91.36 for the control group (for Reasoning IQ it was 84.61 compared with 71.04, respectively; and for Verbal IQ 102.35 and 102.62)—Rosenthal and Jacobson (p. 150) had maintained that this was irrelevant because of the negative correlation between pretest IQ and gain score for Total IQ. According to Elashoff and Snow, however, that correlation was an incorrect statistic for this determination.

The list of criticisms was long. The number of experimental children (bloomers) in each class differed (though this was done purposely, remember, to make it more plausible to the teachers), which affected randomization, and the small samples in some classes precluded analysis within many classrooms. The *p*-values were incorrectly used as measures of strength of effect, and furthermore, in view of the large differences in variances in many comparisons, the "*p*-values quoted ... for comparison in the lower grades are probably spuriously low" (p. 38).

The inadequacy of the TOGA for this particular sample came in for additional censure. To illustrate the extraordinary instability of many scores over the four testing session, 11 additional examples of the scores of individual children were given, a few of which were, for Total IQ: 55, 102, 95, 104; 84, 120, 107, 105; for Verbal IQ: 54, 121, 101, 74; 125, 87, 100, 127; for Reasoning IQ: 0, 77, 82, 143; 114, 81, 88, 106.[11] Histograms and scatterplots revealed the striking deviations from psychometric standards of the score distributions in many of the grades, as did the presence of a number of extreme scores and "outliers" beyond the TOGA's norming range. For example, adjusting for one boy's basic post-test IQ of 202 radically changed the slope of the experimental group's regression line, an important consideration because comparison of pre- to post-test gain scores "will be misleading when their regression slopes are not unity ... or the pretest score distributions are different in the two groups" (p. 97). There was no independent proof of intellectual growth, and in this regard there were no group differences in classroom track transfers (e.g., to a higher track) that would be expected based on the claim that expectancy raised intelligence. Rosenthal and Jacobson "frequently used terms like 'intellectual growth' and 'expectancy advantage' in referring to their dependent variable, never discussing the possibility that their simple IQ gain score might not represent the construct of interest to them" (p. 45).

From their reanalysis of the data, Elashoff and Snow concluded that although results for first and second graders were promising, the pretest differences between experimental and control groups precluded any clear conclusion.[12] "There is enough suggestion of an expectancy effect in grades 1 and 2 to warrant further research, but the RJ experiment certainly does not demonstrate the existence of an expectancy effect or indicate what its size may be" (p. 44). They then made eight recommendations for future research.

In the first sentence of their reply, Rosenthal and Rubin (1971) denied that Elashoff and Snow's critique and reanalysis "impugns the validity of the RJ experiment" (p. 139). Indeed, they wrote, the reanalysis actually supported their conclusions and increased the generality of their results. Specific points and recommendations made by Elashoff and Snow were challenged: A stepwise regression would not have changed the results; rigid null hypothesis decision procedures are not universally recommended; transformations of the data would have statistically biased them; there were no significant differences between experimental and control groups at pretest, so randomization had been successful; and there are indeed increasing effects from higher to lower grade, as shown in a newly presented table and figure. They also presented evidence that the TOGA was a valid instrument for their sample. There was, they argued, not only sufficient peer review in prior publications, but their research was solicited for inclusion in a book prepared for division 9 of the American Psychological Association (APA), was reprinted in other books, and received the first prize of APA's division 13 Cattell Fund Award in 1967.

They protested that Elashoff and Snow had cited only one study by name (the failure of Claiborn 1968 to replicate *Pygmalion*) whereas "numbers of studies showing significant positive effects of teacher expectation had been published and/or read at conventions" (p. 150). They presented a table with percentages of the latest studies (unnamed) of interpersonal expectation reaching significance at various *p*-values, then specifically named four successful teacher expectancy studies that had appeared at the time of the Claiborn study.

This response was an early example of the way that failed post-*Pygmalion* studies of teacher expectancy effects on IQ were either obscured by studies of expectancy effects on other variables or simply disregarded. As Elashoff and Snow (1971) pointed out in their appended rebuttal, none of the four studies assessed the effect of expectancy on *intelligence* (in fact one study measured the effect of expectancy on learning to swim). On the other hand not one of nine studies that did measure teacher expectancy effects on IQ, reviewed by Baker and Crist (1971), succeeded in replicating *Pygmalion*. (Note, however, that Baker and Crist had not included the dissertation of Maxwell 1970.) Rosenthal and Rubin had even failed to mention Rosenthal's own immediate follow-up studies that *did* assess the effects of expectancy on IQ (Anderson & Rosenthal, 1968; Evans & Rosenthal, 1969; Conn et al. 1968), with negative or reverse results.

Elashoff and Snow also maintained (p. 156) that when Rosenthal and Rubin (1971), in their reaffirmation of *Pygmalion*, gave the percentage of all classrooms whose mean IQs showed an advantage (ignoring the size of that advantage) for the experimental children on the basic post-test—76% for Total IQ, 61% for Verbal IQ and 76% for Reasoning IQ—they should also have considered what the percentages already were at pretest—65%, 61% and 76%, respectively.[13] They reiterated questions about randomization. They had recommended stepwise regression only as a way of examining the magnitude of treatment effects. They had not advocated rigid null hypothesis decision procedures, but had urged researchers to make no interpretations of any relationship less than a predetermined *p* value, such as 0.05. Claiming an increasing expectancy advantage when results were dichotomous gave readers the false impression that there were some "positive effects in the middle and higher grades" (p. 160). Extreme scores should have been questioned and handled in some way. The low reliability of the TOGA Reasoning IQs in grades 1 and 2 was worrisome, but a primary concern was the presence of extreme scores and score instability across the four testings. Despite the publications, reprintings and awards, they wrote, "we retain our view that Pygmalion was inadequately and prematurely reported to the general public" (p. 161).

In closing, Elashoff and Snow summarized the points Rosenthal and Rubin ignored, including what they considered perhaps *Pygmalion*'s most basic problem: "[ignoring] the psychological meaning of the scores on which it rests ... What, after all, does an IQ of zero, or 17, or 31, or 202, or 210 really mean? What does an IQ gain of 100, 125, or 135 really mean?" (p. 161).

## Further Studies

The deluge of dissertations that measured the effects of expectancies on intelligence test scores had subsided when Fine (1972) added his to the stock. It differed from the others, as well as from *Pygmalion*, in raising teachers' expectations of student performance on a specific school subject, reading. His sample was drawn from 18 second grade classes at five schools in a low socioeconomic urban area. After excluding the Spanish-speaking students and, for various reasons, a number of other students, 80 experimental and 79 control children contributed to the final IQ data. About 2/3 of the participants were African–American and the remainder non-Hispanic Caucasian. Of the study's 19 teachers, 10 were African–American and 9 non-Hispanic Caucasian.

The teachers were requested to administer an achievement test as well as the group Cognitive Abilities Test (a downward extension of the Lorge–Thorndike Intelligence Tests) which, they were told, was considered "an excellent predictor of future reading achievement" (p. 44). The appropriate teachers were then given a list of those pupils (the experimental group) who on the Cognitive Abilities Test had (allegedly) achieved significantly higher scores than their reading achievement test scores would lead one to expect. The teachers were also told in what months additional requests for assessment of the pupils' reading achievement would be made.

The post-test administration of the Cognitive Abilities Test at the end of the term indicated that raising teachers expectancy of children's reading performance had no effect on IQ (or, for that matter, on reading): the experimental group gained 7.4 IQ points and the control group 9.6 IQ points. Children the teachers remembered (a week after post-test) as children who were expected to show significant progress in reading, and whose teachers also believed in the predictive ability of the Cognitive Abilities Test, did not gain more in IQ than did experimental children of teachers who correctly recalled the children but did not believe in the test. Concerning the unequivocal negative results, Fine implicated, among other things, teachers' familiarity with the students for 4 months prior to the expectancy induction.

A study by Pellegrini and Hicks (1972) has sometimes been cited as showing the effects of expectancy on intelligence (e.g., Raudenbush, 1984), whereas the study's major contribution was its support for the well-known dictum that very often it is "teaching to the test," not a change in general intelligence, that raises IQ. The authors reminded their readers that in *Pygmalion* the teachers did the testing, prior to which they were given information about the rather imposing scientific study in which they were participating. This may have induced them to "familiarize themselves with the criterion measures during the pre-test . . . and given them a feeling of very personal responsibility for the fulfillment of individual prophecies" (p. 414). To test these possibilities, Pellegrini and Hicks made use of the presence of individual child–tutor pairs operating under a county project in which children, recommended for individual instruction, were tutored at least 2 h weekly by volunteer college students. Almost all the children were from low income families and between 70% and 80% were Mexican–Americans. Forty-four elementary school pupils and their tutors participated in the study during the fall 1969 term. Two measures of intelligence, the Peabody Picture Vocabulary Test and the Similarities subtest of the Wechsler Intelligence Scale for Children were used as pre- and post-test measures before and after the 17-week experimental period. Testers were paid assistants unaware of the purpose of the testing.

There were four conditions to which pupil–tutor pairs were randomly assigned. In the high expectation condition the tutors were falsely informed that their pupils had very high intelligence levels, with IQs between 120 and 129, and therefore should make rather dramatic gains in academic areas. A second group of tutors was also given high expectations, but in addition were *familiarized with the test materials*. In the third group the pupils were said to be between 95 and 105 IQ, and in the fourth group the pupils were said to be below average, in the 85 to 95 range. The tutors in the third and fourth groups were told to expect their children to work at those levels and nothing was said about familiarization with the material or academic spurting. In reality, pupils' assignment to tutors was random, "with the restriction that age be approximately equated across groups" (p. 415).

Results can be given succinctly. The children in the high expectancy plus test familiarity condition gained reliably more in Peabody IQ than did any of the other groups, which did not differ from each other. Only when the tutors were familiar with the test material—and therefore could teach it to their pupils (whether they did is not documented)—was there an effect on IQ. High expectation alone was insufficient, failing to confirm *Pygmalion*.

Despite the negative results of his own post-*Pygmalion* studies and those of nearly all others that tried to raise IQ by raising teacher expectancies, Rosenthal defended *Pygmalion* in a 1973 article in *Psychology Today*. This is an instructive document for observing the manner in which teacher expectancy effects on intelligence are engulfed by other expectancy studies, so that a reader unfamiliar with the literature could not possibly suspect the extent to which studies specifically designed to test the expectancy–intelligence relationship were unsuccessful. For example, after commenting that Elashoff and Snow "could not disprove the fact that the experimental children [in *Pygmalion*] did gain more IQ points than did the control children," (p. 59), Rosenthal continued with a discussion not of *Pygmalion* but of the "Pygmalion effect," thereby switching to *all* expectancy studies, as Rosenthal and Rubin (1971) had done in their response to Elashoff and Snow (1971). Referring to the "Pygmalion effect," he then pointed out that of 242 studies "84 found that prophecies, i.e. the experimenters' or teachers' expectations, made a significant difference" (p. 59). The specific effect of teachers' expectations on intelligence, the very essence of Pygmalion, had now been lost, blended not only with teacher expectancy effects on variables other than intelligence, but also with the very general experimenter expectancy effects.

The article is also instructive for the example it provides of how a favored position can be embellished. Concerning *Pygmalion*, Rosenthal (1973) wrote that "teachers had all sorts of good things to say about the 'intellectual bloomers': they had a better chance of being successful in the future, said the teachers; they were more appealing, better adjusted, more affectionate and autonomous" (p. 62). The impression fostered by this description was that teachers were bursting forth with spontaneous comments. In fact, however, they were asked to rate each child, on a scale of 1 to 9, on nine kinds of classroom behaviors (Rosenthal & Jacobson, 1968a, pp. 108–109). Although the ratings of the experimental children differed significantly from the controls in the category of Future Success, there were no statistically significant differences in the categories of Appealing, Adjusted, Affectionate, or Needs Approval.

A year later Rosenthal et al. (1974), citing *Pygmalion* and noting that "studies have shown that the expectation of the classroom teacher can be a significant determinant of her pupils' responses" (p. 115) (replacing "intelligence" with "responses"), introduced a new study in which children in grades 1 to 6 of a "predominantly black inner-city school" (p. 115) were the participants. "Creative potential" was substituted for *Pygmalion's* "academic blooming." The teachers, all of whom were African–Americans, were told that the investigators were developing a measure that would predict "which of their pupils were likely to show greater gains in creativity in the near future" (p. 116). Once again the test was actually the disguised TOGA. In addition, the pupils were asked to draw a picture of a person and "as many different things as possible" (p. 116). Pretests and, after a year, post-tests of both the TOGA and the drawings (later rated for "creativity" by a panel) were obtained. The random selection of 20% of the students and the informing of the teachers were the same as in *Pygmalion*.

Results indicated that for all classes combined there was no difference in TOGA Total IQ, although when grades were tested separately the experimental children in the fifth grade gained reliably more than did their controls. These results differed from *Pygmalion* not only in showing no overall differences in Total IQ but in the appearance of IQ changes in the fifth grade rather than in grades 1 and 2. Did changing teacher expectations from "academic bloomers" to "creative bloomers" affect the results, or was it some other variable? The authors were puzzled: "Why we should obtain significant effects only for the fifth graders is hard to explain as is the failure to find an overall effect" (pp. 119–120), and then noted that differences in teachers, pupils, and "a dependent variable of gain in creativity rather than gain in IQ" (p. 120) might be the source of the different results.

Also in 1974, Sutherland and Goldschmid used six classes of grades 1 and 2 at three schools in middle-class districts in Montreal to test the effects of naturally occurring teacher expectancy. After a school month teachers were asked to rank each pupil's academic potential. The pupils were then given four subtests of the Wechsler Intelligence Scale for Children individually administered by trained experimenters blind as to expectancy rank. Additionally, the regular teachers administered a group test, the Lorge–Thorndike Intelligence Test. Posttests were given at the end of the 5-month experimental period.

The authors divided their sample of 93 pupils into five groups according to level of teacher expectation and found "no significant difference in IQ gain correlating with teacher expectation" (p. 853); that is, all groups rose in mean IQ from pre-to post-test, but not differentially. However, teacher expectancy significantly limited or reversed the mean IQ of a subsample of superior IQ pupils whom teachers inaccurately rated as having only average academic potential, when compared to the pre- to post-test change in performance of pupils whom teachers accurately rated as having above average or superior academic potential. This led the authors to suggest that "changes in IQ score … can be adversely affected when a teacher expects less from a superior student than he is potentially capable of delivering" (p. 854). In the result most relevant for *Pygmalion*, the mean IQ of pupils who had below average IQ and were nevertheless rated as of average academic potential did not rise more than the mean IQs of those accurately rated as having below average or poor academic potential. The authors read their results as confirming the concern of *Pygmalion* critics who were calling for "isolation of the conditions under which expectancy results may operate" (p. 854).

Thus ended, perhaps forever, the empirical studies testing the premise that positive teacher expectancies can raise IQ. The studies are summarized in Table 1.

## Reviews, Discussions and Disputes Continue: 1975–1995

In an article reviewing 50 years of public controversy over mental testing, prepared for an American Academy of Arts and Sciences study, Cronbach (1975a) included a brief description of *Pygmalion* and the public's ignorance of the controversy it had aroused. The public, he wrote, was told "nothing about the controversy but heard much about the study as evidence that mental tests are doing harm" (p. 7). Rosenthal (1975) reacted strongly to Cronbach's statement that *Pygmalion* "merits no consideration as research" (Cronbach, 1975a, p. 6). Cronbach, he wrote, could not be a dispassionate commentator because he had been a principal in the controversy and had written to the publisher that the book's shortcomings would be demonstrated "when his young colleagues Richard Snow and Janet Elashoff at his Stanford Center for Research and Development in Teaching finished their reanalysis of the *Pygmalion* data" (p. 937). Rosenthal illustrated why Elashoff and Snow's reanalysis had failed to discredit *Pygmalion*, and repeated that the *Pygmalion* research had received the Cattell Fund Award. Furthermore, he wrote, Cronbach's implication that the manipulation of teacher belief was too casual to produce such results must be considered in the context of the many experimenter effects that were influenced by just such small treatments. Finally, the "significant" reverse results in the Massachusetts experiment (Conn et al. 1968) cited by Cronbach occurred a year after the students had left the classrooms of the teachers primed for expectancy.

Cronbach (1975b)—who replied that he was not neutral (dispassionate) on this issue nor had he been a principal in the controversy—described the events differently than did Rosenthal. When Rosenthal and Jacobson's book reached him in Tokyo he had prepared a seminar on it. He had also deposited in his files a 1500-word memo that began: "This book taken as a whole is a masterpiece of confusion, aligned with a hypothesis that most of the data contradict" (p. 939). A week later he sent a copy of the memo to Snow. When he returned home he wrote to the publisher, continuing a discussion they had been having "regarding quality standards for semipopular books based on research" (p. 939). He objected to having this private letter to a third party quoted by Rosenthal without permission.

By 1980 there had been enough publications of teacher expectancy effects on IQ that "scorecard" summaries could be given. Jensen (1980) cited 13 studies other than *Pygmalion* that failed to find any such effects. However, one of these (Deitz & Purkey, 1969) did not measure pupils' performance, and two (Pitt, 1956; Dusek & O'Connell, 1973) did not assess the effects of teacher expectancy on IQ. On the other hand he did not cite the dissertation of Maxwell (1970), in which positive expectancy effects on IQ were reported. Still, Jensen's skepticism about optimistic claims for teacher expectancy effects on IQ was not without foundation, as indicated by the results of the concurrent meta-analytic review by Smith (1980) of 47 studies of teacher expectancy effects. Because many of the studies tested more than one effect there was a total of 149 effects, which Smith wisely partitioned into five different categories for separate analyses. The category of "pupil intellectual ability" included 22 effects and produced an average effect size of 0.16, by far the smallest of the five categories. For example, the average effect size for achievement category, also derived from 22 effects, was 0.38. She concluded that the effect of teacher expectancy on intellectual ability was minimal.[14]

Ten years after the latest empirical study, Raudenbush (1984) performed a meta-analysis of the effects of teacher expectancy on IQ. From 18 studies, including *Pygmalion*, he derived 19 effects, 8 of which were negative.[15] He found a small mean effect size, in standard deviation units, of 0.11. Additionally, Raudenbush had replaced the original statistical results of Kester (1969) with a reanalysis (see my footnote 9). However, this inclusive meta-analysis was not Raudenbush's principal interest. Readers will recall how often experimenters had commented that if teachers had spent time with the children they very likely formed an opinion about them before being given false information by the experimenter. In support of the hypothesis that the greater the teachers' familiarity with the pupils the less the potency of the induced expectancy effect (which he believed was consistent with cognitive dissonance theory), Raudenbush produced a scatterplot showing a curvilinear relationship between effect size and weeks of prior teacher–pupil contact. There were many positive effect sizes for studies having limited teacher contact prior to expectancy induction whereas seven of the eight negative effect sizes occurred in studies in which there was more than 2 weeks of prior teacher–pupil contact.

Raudenbush partitioned the studies for further analysis, and presented not only effect sizes but also correlations.[16] In four studies—Flowers (1966), Kester (1969), Carter (1970), and Pellegrini and Hicks (1972)—in which there was no prior pupil–teacher contact before teachers (part-time tutors, in the Pellegrini and Hicks study) were given the expectation or provided with false information, the mean effect size was 0.32 ($r$ = 16). In three studies—Rosenthal and Jacobson (1968a), Keshock (1970), and Maxwell, (1970)—in which there was 1 week of prior contact, mean effect size was 0.26 ($r$ = 0.13). For three studies—Henrikson (1970), Fleming and Anttonen (1971b), and Rosenthal et al. (1974)—with two weeks of prior contact, mean effect size was 0.08 ($r$ = 0.04). For eight studies—Claiborn (1968), Conn et al. (1968), Evans and Rosenthal (1969), Ginsburg (1970), Grieger (1970), Fielder et al. (1971), José and Cody (1971), and Fine (1972)—having more than 2 weeks of prior contact, mean effect size was −0.04 ($r$ = −0.02). Statistical tests of the effects of expectancy on IQ were significant for combined studies in which there were 2 weeks or less of prior contact. Additionally, "expectancy effects were significantly greater than zero only for low-contact studies at grades 1–2 and 7" (p. 93).

In 1985, *Teacher Expectancies*, edited by Dusek (1985), was published, with chapters covering many aspects of teacher expectancy effects but no separate chapter or section on the effects of teacher expectancy on intelligence. In fact, neither intelligence nor IQ appeared in the Index. In a small section of their chapter, Mitlan and Snow reiterated a number of previous criticisms of *Pygmalion*. As part of his chapter, Rosenthal repeated, and updated, his replies to the critics. In the recent criticism by Jensen (1980), he wrote, Jensen had accused him of including studies that used achievement, not intelligence, as an outcome measure. Yet, said Rosenthal, that is exactly what Jensen himself had done. He also chided Jensen for writing that 6.4% of the variance had little practical importance, whereas it is "equivalent to increasing the success rate of a new treatment procedure from 37% to 63%, a change that can hardly be considered trivial" (Rosenthal, 1985, p. 49). Rosenthal then summarized his earlier responses (Rosenthal, 1969a) to Thorndike (1968) and the response Rosenthal and Rubin (1971) had given to Elashoff and Snow (1971), adding two self-criticisms of *Pygmalion*—the questionable use of omnibus *F* tests and the failure to use effect sizes. He said nothing about the bizarre nature of the children's IQs.

Only in the chapter by Hall and Merkel (1985) was there any discussion of the failure to replicate *Pygmalion*, followed by a reaffirmation of the stability of intelligence. Head Start's failure "to permanently increase intelligence test scores is generally well-known," they remarked, and consequently, "if it now turned out that teachers in elementary schools could raise intelligence test scores of randomly selected middle-class students at will, something would clearly be amiss" (p. 71).

Two years later, Wineburg (1987a) critically reviewed the empirical status of the self-fulfilling prophecy "as applied to the effect of teacher expectations on student IQ" (p. 28). Of all the critiques of *Pygmalion*, his provided the most thorough background material as well as the most complete description of the educational, cultural and social climate of the 1950s and 1960s that so readily accommodated the concept of the self-fulfilling prophecy and credulously appropriated *Pygmalion*, though with unpardonable misunderstanding. "Obscured and long forgotten," wrote Wineburg (1987a), "the heart of the Pygmalion controversy was the bold claim that intelligence was affected by teacher expectations" (p. 34). It was this that attracted the media and "influenced the Los Angeles School Board to ban IQ testing in the elementary grades" (p. 34), despite the failure of follow-up studies. Some years earlier, Miller (1980)—although he failed to separate expectancy–intelligence relationships from other expectancy effects—had placed the ready acceptance of the "Pygmalion Effect" and many other beliefs within an academic and societal climate dominated by radical environmentalism. For many (perhaps most) people in the United States, convinced that we would all be of equal intelligence were it not for disparities in environmental circumstances, the findings reported in *Pygmalion* were not (are not) surprising.

Wineburg's article drew a response from Rist (1987) and Rosenthal (1987), whose paper describing the effects of teacher bias on the failure of minority children (Rist, 1970) had been included in Wineburg's critique. Rosenthal (1987) repeated his previous response to critics (see Rosenthal, 1985) and then, to bolster his case, described the meta-analysis performed by Raudenbush (1984). Despite the small amount of variance accounted for, Raudenbush's "effect size estimates suffer from a common problem, the tendency to underestimate the practical importance of behavioral or biomedical interventions" (Rosenthal, 1987, p. 39). Rosenthal's remedy was the "binomial effect size display" (BESD) that he and Rubin had introduced to provide a measure of the practical effect of a new treatment (Rosenthal & Rubin, 1982). As an example, with an effect size $r$ of 0.14, the treatment's success rate would be 57% ($0.50 + r/2$) whereas the control group's success rate would be 43% ($0.50 - r/2$). Raudenbush (1984) had reported BESDs for 0, 1, 2, and >2 weeks of prior pupil–teacher contact. In the four studies with no prior contact ($r = 16$), 58% of the experimental and 42% of the control children would be expected to qualify for a higher track, with the differences between and experimental and control group percentages declining with each additional week of prior contact up to 2 weeks. Rosenthal (1987) cited Raudenbush's finding for the three studies where there were 2 weeks of prior contact and an effect size of 0.08, indicating—despite the minuscule $r$ of 0.04—a positive treatment effect on 52% of the experimental group compared with 48% of the controls. There was no positive effect beyond 2 weeks. In sum, Rosenthal argued that even without *Pygmalion*, and even when results account for a very small part of the variance, they nevertheless provided evidence that teacher expectancies have a practical influence on children's intelligence.[17]

In his rejoinder Wineburg (1987b) returned to the *Pygmalion* data and the fact that many of the kindergarten children raised their post-test Reasoning IQs by attempting many more items than on pretest. Although Rosenthal and Jacobson had described them as having grown intellectually—and Rosenthal (1969a) had mentioned that not trying items did not invalidate IQ tests—Wineburg argued that increased responses could just as easily be attributed to "misunderstood test instructions, uncontrolled test administration, selective teacher coaching, teacher encouragement for guessing, or even chance" (p. 43) (see also Thorndike, 1969). In Wineburg's view, *Pygmalion*'s notoriety resulted from enterprising advocacy and the ready receptivity of its message by the public.

Ten years after his first meta-analysis, Raudenbush (1994) reanalyzed the data from teacher expectancy studies to illustrate the advantages of his random effects meta-analysis, designed to take into account the great number of unidentifiable random effects. He derived 19 "estimates" of treatment effects rather than the 19 "true" (fixed) treatment effects he had derived in his 1984 analysis.[18] The estimated effect size was 0.43 for the four studies and five effect sizes (including two from Pellegrini and Hicks, 1972) where there was no prior pupil–teacher contact. For each week of additional contact (up to 2 weeks) the effect size should reduce by 0.17 points.

That same year Rosenthal (1994) reported that there were now 464 studies of interpersonal expectancy effects, with an overall effect size of 0.63, and he repeated the inspirational seed for *Pygmalion*: If rats can be made brighter when expected to, why not children? Despite this, he did not treat as a separate domain those studies that attempted to make children brighter, except in a footnote, where he mentioned that "Raudenbush found very strong evidence ($r = 0.67$) that substantial effects of teacher expectancies could be demonstrated only when the induction of the expectancies was credible (i.e., when teachers had known pupils only 2 weeks or less at the time they were given the expectation)" (p. 170). However, this specific correlation had been given not by Raudenbush but by Rosenthal (1987) himself, who, following Raudenbush, used a median test in which the dividing point was studies where teacher–pupil contact was 2 weeks or less. Raudenbush (1984) had reported an even higher $r$ of $-0.77$ for 18 studies when the curvilinearity of the relationship between effect size and period of prior contact was taken into account. The high correlation tells us that there was a substantial negative relationship between effect size and prior contact but says nothing about the strength of the effect sizes.

In his comments on Rosenthal's latest report, Snow (1995) alluded once again to the failure of Rosenthal (1994) to treat the expectancy–intelligence studies separately. Rosenthal, he said, had cloaked the lack of support for *Pygmalion* "in meta-analyses of related though not comparable research" (p. 169). Snow then plotted *Pygmalion*'s pretest Reasoning IQs against the post-test Reasoning IQs and placed a square outline around all IQs between 60 and 160 (TOGA's norm range, about 65% of the scores), thereby revealing that the "expectancy effect disappears when extreme scores [which had to be extrapolated] were omitted. The heightened experimental regression line . . . in the total group appears to result solely from five children whose respective pretest–post-test scores were 17–110, 18–122, 133–202, 111–208, and 113–211" (p. 170). As for meta-analysis of Raudenbush (1984), Snow (1995) reversed the abscissa and ordinate of Raudenbush's figure and drew a line at the zero effect size point, thereby showing more clearly the eight negative effect sizes. He inserted arrows to identify the Pellegrini and Hicks (1972) and *Pygmalion* data points, and also pointed out that because of the skewness of the data from the 18 studies the median effect size (0.035) was more appropriate than the mean effect size (0.11) given by Raudenbush.

Rosenthal (1995) replied that even after omitting *Pygmalion* the median effect size for the nine studies with 2 weeks or less of teacher–pupil contact was 0.18, and he cited Raudenbush's (1994) latest meta-analysis of studies having no prior teacher contact. He pointed out once again (this time in a footnote) that *Pygmalion* was awarded First Prize of the Cattell Fund Award of APA's Division 13, and closed with resolute finality: "Even if Lenore Jacobson and I had never conducted our experiment, there are now too many studies for even committed criticisms of disliked results to make the basic conclusion go away: Teachers' expectations can affect pupils' intellectual functioning. Science is the loser when new data have no effect on prior belief" (p. 172). Thus ended, for the time being, the published exchanges that characterized the controversy over *Pygmalion* and its claim that teacher expectancy can raise intelligence.

## Final Comments

Many of those who are most critical of *Pygmalion*'s claim that teacher expectancies can raise intelligence have nevertheless expressed their belief that expectancy influences many other kinds of performance.

> In closing, let me express a very real interest in the notion of the "self-fulfilling prophecy." I would expect the phenomenon to appear most clearly ... in those areas that are most directly teacher-based and school-dependent, such as learning to read, to write and to cipher. Perhaps others can learn from *Pygmalion*'s shortcomings, and carry out research on these problems that is psychometrically and experimentally adequate. (Thorndike, 1969, p. 692)

> Within education, the issue has never been whether teachers form expectancies or whether these expectancies affect students in sundry subtle and not-so-subtle ways . . . . But regarding the dispute that has come to be known as the "*Pygmalion* controversy" such questions missed the point. (Wineburg, 1987a, p. 34)

> I agree that general evidence shows that interpersonal expectancies exist as psychological phenomena, and that teacher expectancies, as an example, can influence classroom teaching and learning, at least sometimes. (Snow, 1995, p. 169)

*Pygmalion*'s central thesis was that raising teacher expectations will raise students' general intelligence. This thesis was purportedly supported by *Pygmalion*'s results in which, on average, the entire experimental group outgained the entire control group, as measured by Total IQ. However, based on the "quasireplications" in which Rosenthal participated, as well as many other studies, the *Pygmalion* findings have been uncommonly difficult to replicate, a fact that was dissipated in subsequent reviews by pooling expectancy–intelligence studies with other kinds of expectancy studies. Examples of this proclivity have been cited throughout this review and many others could be added (e.g., Cooper, 1979; Jones, 1986). Early statements that teacher expectancy effects on intelligence were exceedingly complex and, indeed, not universal (Rosenthal, 1969b; Rosenthal and Jacobson, 1968a; Conn et al., 1968) had raised the prospect that Rosenthal and others would meet the problems forthrightly, a prospect that was never fulfilled. Indeed, despite *Pygmalion*'s incomprehensible IQ data, defenders and many reviewers believed (believe) there were no flaws so inordinate as to invalidate the original experiment. In many studies, interactions and unexpected findings for various subgroups on different subscales were submitted as supportive evidence for teacher expectancy effects on intelligence. This tactic frequently drew comments that unpredicted findings, when produced by post hoc digging into subsamples, should only be used as a prediction for new experiments, not as synthetic confirmation of the original experimental prediction.[19]

If we follow expectancy theory to its logical conclusion the possibility must be entertained that those who study expectancy effects are at the same time exhibiting them. This was reflected in the title of the review by Buckley (1968) and has been mentioned so often that it hardly needs repeating (e.g., Barber, 1978; Gadin, 1978). Even the study of studies on expectancy effects must be plagued by expectancy effects, and so on in infinite regress (Jung, 1978). As a possible remedy, Rosenthal and Rubin (1978b, pp. 412–413) suggested that expectancy investigators be randomly divided into two groups, after which the expectancy of the principal investigators of one group be withheld from the experimenters. The belief that a remedy is needed implies a belief in expectancy effects, even on so basic a trait as intelligence, but why then has *Pygmalion* been so difficult to replicate? Is it because many experimenters expected negative results? The permutations are staggering.

This review finds no compelling evidence for a clearly defined method that will reliably raise students' intelligence by manipulating teachers' expectations. The data that went into Raudenbush's meta-analyses do not seem to me to provide a strong enough basis for the blanket assertion of Rosenthal (1995) that teacher expectations can effect pupils' intelligence, though they highlight definitive issues for anyone interested in still further experiments.

Rosenthal (1985) had written that the bulk of the criticism of *Pygmalion* came from the field of educational psychology, not from mathematical statisticians, social psychologists or educators—except for the president of the teachers' union (see Shanker, 1971). "We leave [this] observation as just a curiosity," he wrote, "one that might be clarified by workers in the fields of the history, sociology, and psychology of science" (Rosenthal, 1985, p. 49). I believe *Pygmalion* and its sequelae might be of interest to them, but only as an illustration of how objectivity, caution and skepticism have too often been replaced by promotion and advocacy. The other lesson this history teaches is one that never seems to be learned: that unexamined, premature enthusiasm for newly minted, seemingly effortless psychological and pedagogical methods for curing intractable disorders and raising intelligence are chimeras that lead only to disillusion (Spitz, 1986, 1997).

## NOTES

1. Effects of expectancy on many variables other than intelligence have been reported in numerous experiments with nonhuman and human subjects, carried out not only in psychological laboratories but also in factories, clinics and offices. Here, *Pygmalion* and studies of the effects of expectancy on human intelligence will not be merged with other studies, as they often have been in the past (e.g., Rosenthal & Rubin, 1971, 1978a; Rosenthal, 1994).

2. Unless otherwise noted, the descriptions that follow are drawn from the book.

3. Some writers have used the description sometimes given by Rosenthal and Jacobson (1968a) that the Harvard Test "was represented to the teachers as one that would predict *intellectual* 'blooming' or 'spurting'"

(p. 175, emphasis added), which is what they wrote in two places (including the Abstract) of a previous paper (Rosenthal & Jacobson, 1966). Elsewhere in the 1966 paper they said they disguised the test "as a test designed to predict academic 'blooming' or intellectual gain" (p. 115), which is what they wrote in still another publication (Rosenthal & Jacobson, 1968b, p. 21). But elsewhere in their book (Rosenthal & Jacobson, 1968a), they wrote that the test was "purported [to the teachers] to be a predicator of academic 'blooming' or 'spurting'" (p. 66), with no mention of intellectual blooming, correctly portraying the description in the information sheet given to each teacher and reprinted on page 66 of their book.

4. The possible environmental contributions have yet to be validated, but findings suggest that to a surprisingly large extent children create their own environments (e.g., Scarr, 1996).

5. In the abstract of the first journal article to formally present their results, Rosenthal and Jacobson (1966) wrote that the effects of "teachers' expectancies operated *primarily* among the younger children" (p. 115, emphasis added). In actuality, statistically significant results were found *only* among the younger children. Note that, for the most part, one-tailed (directional) tests were used in this study even though unpredicted results were also reported. Rosenthal and Jacobson (1968a) remarked that although not strictly in accord with the logic of using one-sided tests (as they did when predicting the direction of differences), two-tailed (nondirectional) tests were used for unexpected results "as an aid to those who would prefer the use of two-tailed tests throughout and who will have to double all $p$'s given as one-tail" (p. 95). Questions concerning the use of one-tailed tests were subsequently raised by critics.

6. For the recognition test, teachers were given sets of four names and asked to indicate which of the four had been designated as special children at the beginning of the preceding academic year. At each presentation two of the four names were experimental and two were control children. The score for each set was the number of experimental children correctly identified minus the number of control children misidentified as experimental children, and could range from $-2$ (two false positives) to $+2$ (two correct choices). The mean recognition score above chance was $+0.44$, $p < 0.04$, using a one-tailed test (Rosenthal & Jacobson, 1968a, p. 155), which was of marginal significance using a two-tailed test. The correlation of recognition scores and magnitude of expectancy effects was not statistically reliable.

7. This full report was published 2 years later in *Pygmalion Reconsidered* by Elashoff and Snow (1971) and we will turn to it shortly. Janet Elashoff, Snow's Stanford University colleague, had assisted Snow in his original review.

8. Children with mental retardation have also participated in experiments on the effects of expectancy on such non-IQ variables as academic performance and social development (Gozali & Meyen, 1970; Haskett, as cited in Gozali & Meyen, 1970).

9. Later, Raudenbush (1984) reported in a footnote that a reanalysis of data of Kester (1969) using an analysis of covariance revealed an expectancy effect size of 0.27 and a one-tailed $p$-value of 0.05.

10. Both Fleming and Anttonen were on the doctoral committee of Keshock (1970), and Fleming was on that of Maxwell (1970).

11. Nevertheless the test–retest correlations from pretest to the basic post-test (from first to third testing) for the combined first and second grade control and experimental groups on Total IQ were, respectively, 0.66 and 0.72, so that despite some extraordinary score changes the children's ranking remained fairly consistent. For Reasoning IQ, where very large gains were made by both groups, the correlations were 0.45 and 0.50 (Elashoff & Snow, 1971, Table 6).

12. Elashoff and Snow emphasized that the only real test of the claim that teacher expectancy can raise intelligence is replication with more reliable data: "Definitive conclusions require additional experiments" (Elashoff & Snow, 1971, p. 124).

13. The percentages I calculated here differ somewhat from those of Elashoff and Snow (1971, p. 157), apparently because of a difference in addition. For the percentage of classrooms in which experimental children *gained* more in mean IQ than control children at post-test, only in Reasoning IQ was there a significant effect: more gain in 15 of 17 classrooms (88%, $p < 0.05$ two-tailed). For neither Total IQ gain nor Verbal IQ gain did percentage of classrooms show a significant experimental effect. Note that many classrooms had three or fewer experimental children; two classrooms had only 1 and another only 2 (range of 1 to 8, median of 4) at the post-test. On the other hand, the number of control children ranged from 10 to 20 per classroom with a median of 15.

14. Presumably she included *Pygmalion* in the intellectual ability category, but unfortunately Smith's reference section was available only on request and I was unable to contact her.

15. Raudenbush included separately two of the conditions in the Pellegrini and Hicks (1972) study as "tester aware" and "tester blind," meaning testers were aware or not aware of which students were in the high expectancy group. In fact, however, the testers were unaware of student placement in *both* these conditions, and in

both conditions the tutors were given high expectations for their students. It was the *tutors*, not the testers, who were or were not aware, and awareness meant familiarization with the tests. Including these two conditions (effect sizes 0.85 for familiarized tutors and 0.19 for unfamiliarized tutors) separately in the meta-analysis produced the 19 effect sizes from the 18 studies. Snow (1995) later questioned Raudenbush's inclusion of the tutor aware (that is, familiarized with the test) condition, which produced the highest effect size in the meta-analysis. Note also that in this and all subsequent analyses Raudenbush excluded the Anderson and Rosenthal (1968) study (where there were no expectancy effects on Total IQ) because it "studied mentally retarded children" (p. 88).

16. The $r$'s can be derived by dividing effect size by the square root of the squared effect size plus 4. Squaring the $r$ then gives the percentage of the variance accounted for. An effect size of 0.32 accounts for about 2.5% of the variance. Raudenbush (1984) had also found a substantial inverse relationship between sample size and effect size ($r = -0.36$), but "the effect of prior contact was largely independent of sample size" (p. 91).

17. It is one thing when a very small effect size indicates an increased probability that a treatment saves lives (dichotomous data), however few, but quite another thing when a very small effect size indicates an increased probability that a procedure raises a few IQs (continuous data) less than the amount of the test's error of measurement. There is a sharp contrast between the permanence of the former and the vagueness and impermanence of the latter. Note also that the BESD is appropriate for continuous data when the variances and sample sizes of the two conditions are similar.

18. The random effects approach has some problems: It treats the variance "as if it were known when in fact it must be estimated from the data," and the assumption is made "that the random effects are normally distributed" (Raudenbush, 1994, p. 316), whereas they may not be.

19. Barber and Silver (1968) found these and related problems to be pervasive in studies of experimenter expectancy effects (as well as teacher expectancy effects). The running controversy that Rosenthal and his colleagues have had with Barber and his colleagues (e.g., Barber, 1978) over the experimenter bias effect parallels their controversy with Snow and his associates over teacher expectancy effects on intelligence.

## 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, J. P., & Crist, J. L. (1971). Teacher expectancies: A review of the literature. In J. D. Elashoff & R. E. Snow (Eds.), *Pygmalion reconsidered: A case study in statistical inference: Reconsideration of the Rosenthal–Jacobson data on teacher expectancy* (pp. 48–64). Worthington, OH: Charles A. Jones.

Barber, T. X. (1978). Expecting expectancy effects: Biased data analyses and failure to exclude alternative interpretations in experimenter expectancy research. *Behavioral and Brain Sciences, 1*, 388–390.

Barber, T. X., & Silver, M. J. (1968). Fact, fiction, and the experimenter bias effect. *Psychological Bulletin Monograph Supplement,* part 2*, 70*(6), 1–29.

Buckley, J. J. (1968). Who is Pygmalion, which is Galatea? [Review of *Pygmalion in the classroom*]. *Phi Delta Kappan, 50*, 124.

Carter, D. L. (1970). The effect of teacher expectations on the self-esteem and academic performance of seventh grade students. *Dissertation Abstracts International* (University Microfilms No. 71-7612)*, 31*(09), 4539A.

Claiborn, W. L. (1968). An investigation of the relationship between teacher expectancy, teacher behavior, and pupil performance. *Dissertation Abstracts* (University Microfilms No. 69-8619)*, 29*, 991A.

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

Coles, R. (1969, April 19). What can you expect? [Review of the book *Pygmalion in the classroom*]. The New Yorker, pp. 169–170, 173–177.

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. (1979). Pygmalion grows up: A model for teacher expectation communication and performance influence. *Review of Educational Research, 49*, 389–410.

Cronbach, L. J. (1975a). Five decades of public controversy over mental testing. *American Psychologist, 30*, 1–14.

Cronbach, L. J. (1975b). Cronbach replies. *American Psychologist, 30*, 938–939.

Deitz, S. M., & Purkey, W. W. (1969). Teacher expectation of performance based on race of student. *Psychological Report, 24*, 694.

Dusek, J. B. (Ed.). 1985. *Teacher expectancies*. Hillsdale, NJ: Erlbaum.

Dusek, J., & O'Connell, E. (1973). Teacher expectancy effects on the achievement test performance of elementary school children. *Journal of Educational Psychology, 65*, 371–377.

Elashoff, J. D., & Snow, R. E. (1971). *Pygmalion reconsidered: A case study in statistical inference: Reconsideration of the Rosenthal–Jacobson data on teacher expectancy.* Worthington, OH: Charles A. Jones.

Elliott, R. (1987). *Litigating intelligence: IQ tests, special education, and social science in the classroom*. Dover, MA: Auburn House.

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

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

Fine, L. (1972). The effects of positive teacher expectancy on the reading achievement and I.Q. gains of pupils in grade two. *Dissertation Abstracts International* (University Microfilms No. 72-27180)*, 33*(04), 1510A.

Fiske, D. W. (1978). The several kinds of generalization. *Behavioral and Brain Sciences, 1*, 393–394.

Flanagan, J. C. (1960). *Tests of general ability: Preliminary technical report*. Chicago: Science Research Associates, Inc.

Fleming, E. S., & Anttonen, R. G. (1971a). Teacher expectancy as related to the academic and personal growth of primary-age children. *Monographs of the Society for Research in Child Development* (5, Serial No. 145)*, 36*.

Fleming, E. S., & Anttonen, R. G. (1971b). Teacher expectancy or My Fair Lady. *American Educational Research Journal, 8*, 241–252.

Flowers, C. (1966). Effects of an arbitrary accelerated placement on the tested academic achievement of educationally disadvantaged students. *Dissertation Abstracts* (University Microfilms No. 66-10288)*, 27*, 991A.

Gadin, H. (1978). Great expectations ... big disappointment. *Behavioral and Brain Sciences, 1*, 394.

Ginsburg, R. E. (1970). An examination of the relationship between teacher expectancies and students' performance on a test of intellectual functioning. *Dissertation Abstracts International* (University Microfilms No. 71-922)*, 31*(07), 3337A.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatures, history, and bibliography. *Intelligence, 24*, 13–23.

Gozali, J., & Meyen, E. L. (1970). The influence of the teacher expectancy phenomenon on the academic performances of educable mentally retarded pupils in special classes. *Journal of Special Education, 4*, 417–424.

Grieger, R. M. II. (1970). The effects of teacher expectancies on the intelligence of students and the behavior of teachers. *Dissertation Abstracts International* (University Microfilms No. 70-26291)*, 31*(07), 3338A.

Hall, V. C., & Merkel, S. P. (1985). Teacher expectancy effects and educational psychology. In J. B. Dusek (Ed.), *Teacher expectancies* (pp. 67–92). Hillsdale, NJ: Erlbaum.

Henrikson, H. A. (1970). An investigation of the influence of teacher expectation upon the intellectual and achievement performance of disadvantaged kindergarten children. *Dissertation Abstracts International* (University Microfilms No. 71-14791)*, 31*(12), 6278A.

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

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

Jones, E. E. (1986). Interpreting interpersonal behavior: The effects of expectancies. *Science, 234*, 41–46.

José, N. J. (1969). Teacher–pupil interaction as it relates to attempted changes in teacher expectancy of academic ability and achievements. *Dissertation Abstracts International* (University Microfilms No. 70-7291)*, 30*(10), 4277A.

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

Jung, J. (1978). Self-negating functions of self-fulfilling prophecies. *Behavioral and Brain Sciences, 1*, 397–398.

Keshock, J. D. (1970). An investigation of the effects of the expectancy phenomenon upon the intelligence, achievement and motivation of inner-city elementary school children. *Dissertation Abstracts International* (University Microfilms No. 71-19010)*, 32*(01), 243A.

Kester, S. W. (1969). The communication of teacher expectations and their effects on the achievement and attitudes of secondary school pupils. *Dissertation Abstracts International* (University Microfilms No. 69-17653)*, 30*, 1434A.

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

Kohl, H. (1968, Sept. 12). Great expectations [Review of *Pygmalion in the classroom*]. *New York Review of Books*, 31–33.

Leo, J. (1967, Aug. 18). Study indicates pupils do well when teacher is told they will. The New York Times, pp. 1, 20.

Maxwell , M. L. (1970). A study of the effects of teacher expectation on the IQ and academic performance of children. *Dissertation Abstracts International* (University Microfilms No. 71-1725)*, 31*(07), 3345A.

Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review, 8*, 193–210.

Miller, H. L. (1980). Hard realities and soft social science. *Public Interest, 59*, 67–82.

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

Pitt, C. C. V. (1956). An experimental study of the effects of teacher's knowledge or incorrect knowledge of Pupil I.Q.'s on teachers' attitudes and practices and pupils' attitudes and achievement. *Dissertations Abstracts* (University Microfilms No. 56-19254)*, 16*(12), 2387–2388.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76*, 85–97.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russel Sage Foundation.

Rist, R. C. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review, 40*, 411– 451.

Rist, R. C. (1987). Do teachers count in the lives of children? *Educational Researcher, 16*(9), 41– 42.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Rosenthal, R. (1969a). Empirical vs. decreed validation of clocks and tests. *American Educational Research Journal, 6*, 689–690.

Rosenthal, R. (1969b). Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 181–277). New York: Academic Press.

Rosenthal, R. (1973). The Pygmalion effect lives. *Psychology Today,* 7(4), 56, 58–60, 62–63.

Rosenthal, R. (1975). On balanced controversy. *American Psychologist, 30*, 937–938.

Rosenthal, R. (1985). From unconscious bias to teacher expectancy effects. In J. B. Dusek (Ed.), *Teacher expectancies* (pp. 37–65). Hillsdale, NJ: Erlbaum.

Rosenthal, R. (1987). Pygmalion effects: Existence, magnitude, and social importance. *Educational Researcher, 16*(9), 37–41.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science, 3*, 176–179.

Rosenthal, R. (1995). Critiquing *Pygmalion*: A 25-year perspective. *Current Directions in Psychological Science, 4*, 171–172.

Rosenthal, R., Baratz, S. 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. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Report, 19*, 115–118.

Rosenthal, R., & Jacobson, L. (1968a). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Holt, Rhinehart & Winston.

Rosenthal, R., & Jacobson, L. F. (1968b). Teacher expectations for the disadvantaged. *Scientific American, 218*(4), 19–23.

Rosenthal, R., & Jacobson, L. (1968c). Self-fulfilling prophecies in the classroom: Teachers' expectations as unintended determinants of pupils' intellectual competence. In M. Deutsch, I. Katz, & A. R. Jensen (Eds.), *Social class, race, and psychological development* (pp. 219–253). New York: Holt, Rinehart and Winston.

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

Rosenthal, R., & Rubin, D. B. (1971). *Pygmalion* reaffirmed. In J. D. Elashoff & R. E. Snow (Eds.), *Pygmalion reconsidered: A case study in statistical inference: Reconsideration of the Rosenthal–Jacobson data on teacher expectancy* (pp. 139–155). Worthington, OH: Charles A. Jones.

Rosenthal, R., & Rubin, D. B. (1978a). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 1*, 377–386.

Rosenthal, R., & Rubin, D. B. (1978b). Issues in summarizing the first 345 studies of interpersonal expectancy effects. *Behavioral and Brain Sciences, 1*, 410–415.

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

Scarr, S. (1996). How people make their own environments: Implications for parents and policy makers. *Psychology, Public Policy and the Law, 2*, 204–228.

Shanker, A. (1971). False research vs. the public schools. *The New York Times,* Section 4, 13.

Smith, M. L. (1980). Teacher expectations. *Evaluation in Education, 4*, 53–55.

Snow, R. E. (1969). Unfinished Pygmalion [Review of *Pygmalion in the classroom*]. *Contemporary Psychology, 14*, 197–199.

Snow, R. E. (1995). Pygmalion and intelligence? *Current Directions in Psychological Science, 4*, 169–171.

Spitz, H. H. (1986). *The raising of intelligence: A selected history of attempts to raise retarded intelligence*. Hillsdale, NJ: Erlbaum.

Spitz, H. H. (1997). *Nonconscious movements: From mystical messages to facilitated communication*. Mahwah, NJ: Erlbaum.

Sutherland, A., & Goldschmid, M. L. (1974). Negative teacher expectation and IQ change in children with superior intellectual potential. *Child Development, 45*, 852–856.

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

Thorndike, R. L. (1969). But you have to know how to tell time. *American Educational Research Journal, 6*, 692.

Wineburg, S. S. (1987a). The self-fulfillment of the self-fulfilling prophecy: A critical appraisal. *Educational Researcher, 16*(9), 28–37.

Wineburg, S. S. (1987b). Does research count in the lives of behavioral scientists? *Educational Researcher, 16*(9), 42–44.

Wolf, E. P. (1981). *Trial and error: The Detroit school segregation case*. Detroit, MI: Wayne State University.

Zuroff, D. C., & Rotter, J. B. (1985). A history of the expectancy concept in psychology. In J. B. Dusek (Ed.), *Teacher expectancies* (pp. 9–36). Hillsdale, NJ: Erlbaum.