The Expectations of Pygmalion's Creators

CHARLOTTE P. TAYLOR*

BEFORE the publication of *Pygmalion in the Classroom* by Robert Rosenthal and Lenore Jacobson, preliminary articles and conference speeches reported to the educational public the authors' view that teachers' expectations appear to affect their pupils' intellectual development.[1]

This major thesis was not unexpected by many educators because of the feeling in a large segment of the professional community. This climate could be characterized as having a perceptual approach to epistemology, a transactional approach to interpersonal relations, and a self-concept approach to developmental process. The Pygmalion study slipped into this intellectual milieu with very little dissonance. Further, it appeared to be buttressed by internal consistency that reflected further expectations of the society at large; for example, teachers' expectations apparently had a differential effect upon minority group children.

In marked contrast to this initial reception was the immediate response of the professional community to another study which seemed to be addressed to a similar issue: "How Much Can We Boost IQ and Scholastic Achievement?"[2] Harvard professor Arthur Jensen's thesis is that measured intellectual differences between races are based on real genetic differences. This study, however, is in radical opposition to prevailing social expectations. In two issues of the *Harvard Educational Review* and bulletin of ERIC—Information Retrieval Center on the Disadvantaged—"the big guns" of the educational establishment massed an attack on Jensen's unpopular hypothesis.[3]

Rosenthal and Jacobson elected to present the full development and analysis of their experiment in a form aimed somewhere between layman and expertly trained experimenter. For example, raw data are not included in *Pygmalion in the Classroom*. Perhaps this is another reason that rigorous evaluations of the study have been few. Nevertheless, three studies appearing in the professional literature[4,5,6] have raised such

[1] Robert Rosenthal and Lenore Jacobson. *Pygmalion in the Classroom*. New York: Holt, Rinehart and Winston, Inc., 1968.

[2] Arthur R. Jensen. "How Much Can We Boost IQ and Scholastic Achievement?" *Harvard Educational Review* 39 (1): 1-124; Winter 1969.

[3] See: *Harvard Educational Review* 39 (2 and 3); also *IRCD Bulletin* 5 (4): 1-24; Fall 1969. New York: ERIC—Information Retrieval Center on the Disadvantaged, Horace-Mann—Lincoln Institute, Teachers College, Columbia University.

[4] William L. Claiborn. "Expectancy Effects in the Classroom: A Failure To Replicate." *Journal of Educational Psychology* 60 (5): 377-83; October 1969.

[5] Peter Gumpert and Carol Gumpert. "The Teacher as Pygmalion: Comments on the Psychology of Expectation." *Urban Review* 3 (1): 21-25; September 1968.

[6] Robert L. Thorndike. Review of Rosenthal and Jacobson, *Pygmalion in the Classroom*. *American Educational Research Journal* 5 (4): 708-11; November 1968.

* *Charlotte P. Taylor, Director of Instruction, Newark School District, Newark, Delaware*

serious questions about the validity of the Pygmalion data and their statistical analysis that the conclusions of the entire study seem in doubt. Those readers of *Educational Leadership* who can find their way about in the sophistication of statistics would do well to study these reports carefully.

Even a cursory review, however, suggests serious scientific impediments to accepting Rosenthal and Jacobson's conclusions.

Failure To Replicate

Central to the scientific method is the criterion that an experiment's reliability is based on the ability of other scientists to repeat it and arrive at the same results. Two replication studies are cited in *Pygmalion in the Classroom*. One conducted by Judy Evans using a similar design in another school resulted in what the authors themselves call a "dramatic and very highly statistically significant reversal." [7]

As supportive replication, Rosenthal and Jacobson offer results from a doctoral dissertation by W. Victor Beaz. In this study, teachers expected good or poor symbol learning from individual Head Start pupils. Those teachers who expected good learning "tried to teach more symbols to their pupils than did the teachers given unfavorable expectations." [8] Change in test behavior was measured on a performance test of symbol learning—the exact skill which teachers did or did not attempt to teach. It is claimed, however, that (unidentified) measures of statistical control removed the influence of this instructional factor and a significant difference in achievement remained.

In an experiment of two months' duration between pre- and post-tests, using 12 first-grade classrooms and the Test of General Ability (TOGA) which Rosenthal and Jacobson had used, Claiborn failed to replicate the findings of the Pygmalion study.[9] An important difference in his procedure may have been the attempt to control initial pretest differences—which, as is shown below, had a basic influence upon the Pygmalion results. Claiborn's article is also worthwhile for highlighting the major errors in statistical analysis which characterize not only the manipulation of the Pygmalion data but previous laboratory expectancy studies conducted by Rosenthal from which the within-school study developed.

Control of Interactions

In 1968, the *Urban Review* paired the key chapter of *Pygmalion*, "The Magic Children of Galatea," with a critical analysis by Carol and Peter Gumpert.[10] While "concurring that the analysis of the data as presented supported the major hypothesis," the Gumperts raised questions about the interactions of factors such as sex which were the focus of many of the Rosenthal and Jacobson subconclusions. For example, ". . . how the strength of teacher expectation varied with the sex of the students, their group status (minority or otherwise), their ability track, and so on—is a complex and confusing business. It is fairly easy to interpret each result taken alone, but the reader who attempts to make his interpretations consistent with one another will be in for a good deal of difficulty. . . ." [11] Basic IQ data, for example, are not presented by sex.

Analyzing the Pygmalion data as given for the effect of individual classroom, the Gumperts found a significant variable unrecognized by Rosenthal and Jacobson. How much influence was caused by classroom atmosphere and interaction, by peer- and self-esteem? Before discussing suggestive factors which may create an expectation bias, the Gumperts also score the Pygmalion study's far too great reliance upon the validity of pretest scores and the authors' confusion of *stability* of statistical significance levels with the *magnitude* of a statistical difference.

Validity of Data

At once the most concise and most decisive criticisms of *Pygmalion in the Class-*

[7] Rosenthal and Jacobson, *op. cit.*, p. 94.

[8] *Ibid.*, p. 182.

[9] Claiborn, *op. cit.*

[10] Gumpert and Gumpert, *op. cit.*

[11] *Ibid.*, p. 21.

room were presented by Robert L. Thorndike in the November 1968 issue of the *American Educational Research Journal.*[12] A year later, Rosenthal responded in the same journal.[13]

Thorndike focused on just one cell of the pretest data—which reported that 19 first-grade children had a mean IQ of 31 on the subtest score of reasoning. Since the children tested were at the bottom of the TOGA age span, extrapolation alone (based on only two correct responses) could suggest an IQ of 31. This score combined with the other classes resulted in a mean reasoning IQ of 58 for the entire first- and second-grade classes.

Rosenthal responded, however, that since Thorndike focused on the reasoning scores, he apparently accepted the total scores as accurate. This is an unlikely and inappropriate rebuttal since the total IQ score on TOGA is a simple addition of the subscores of verbal concepts and reasoning—an error of one guarantees an error in the other.

Further, Rosenthal countered that 15 of 17 classrooms showed the Pygmalion bias on the reasoning IQ—or 13 out of 14 if one omitted the first and second grades. (One of these differences shows statistical significance.)

Thorndike also questioned a mean post-test IQ of 150 for six second-grade "bloomers." This would have demanded that the pupils got *all* questions correct on TOGA. How then could an accompanying standard deviation of 40 suggest some pupils had scored *more* than *all* correct?

Rosenthal's response: "Their mean IQ was simply 1.5 times the magnitude of their mean CA." A simple division using the MA's given by Rosenthal as support of this statement reveals that two of the six children sampled in a *high* track section of second grade were 11 years old or approximately 2½ years *retarded* in grade level. Because of the small size of the sample, such an unusual occurrence has a very powerful effect on the combined scores.

[12] Thorndike, *op. cit.*

[13] Robert Rosenthal. "Empirical vs. Decreed Validation of Clocks and Tests." *American Educational Research Journal* 6 (4): 689-91; November 1969.

For Thorndike, such questionable, extrapolated scores could not be manipulated mathematically with any degree of accuracy. Three times four may be twelve—but if three is a guess, what meaning had twelve? And what meaning a *t*-score for statistical significance that involves a series of manipulations?

Rosenthal claimed that the scores had meaning because they predicted placement in low, medium, or high track first-grade classes.

A simple ordering scale permitting the ranking of items, however, does not possess the accuracy of a scale measured in equal differences. It is this latter which statistical analysis demands. (For example, to devise an IQ, one must divide a numerical MA by a numerical CA. One cannot construct an IQ from the statement, "Sam is smarter than Mary, divided by chronological age.")

A third issue raised by Thorndike dealt with the impact of omission of test items. This probability was offered to explain obtaining correct answers well below chance (for example, 2 out of 28 vs. an expected 5 or 6). Rosenthal countered that the creator of TOGA had tabulated an MA for a *single* correct item . . . all other items might be omitted.

Thorndike's response was that extra encouragement from a biased teacher might well lead a youngster to try a few more items—and this behavior alone would result in a very much higher post-test score.[14] It would be erroneous to construe this as an increase in intellectual development.

The total IQ scores which result from the extraordinary pretest reasoning subtest scores, their use in conjunction with scores across grades, and the gains in scores which grow out of them are the accumulative results upon which depend the majority of the conclusions of the Pygmalion study. If these basic building blocks are invalid, the edifice built upon them totters. Further, as suggested in the other studies, the statistical cement which attempts to bind them together is of faulty composition.

It appears that *Pygmalion in the Classroom* itself reflects the expectancy bias which is the creation of its authors. As Thorndike concluded: "Perhaps others can learn from *Pygmalion*'s shortcomings, and carry out research on these problems that is psychometrically and experimentally adequate."[15] Such efforts by other scientists are certainly needed before the story of Galatea's children may be removed from educational mythology to the realm of scientific fact. □

[14] Robert L. Thorndike. "But You Have To Know How To Tell Time." *American Educational Research Journal* 6 (4): 692; November 1969.

[15] *Ibid.*